Sentiment Analysis of Text Guided by Semantics and Structure

Sentiment Analysis of Text Guided by Semantics and Structure

Sentimentanalyse van tekst geleid door semantiek en structuur

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board

The public defense shall be held on

Friday 13 November 2015 at 13:30 hours

by

Alexander Cornelis Hogenboom
born in Dordrecht, The Netherlands.

Erasmus University Rotterdam

Doctoral Committee

Promotors:	Prof.dr.ir. U. Kaymak
	Prof.dr. F.M.G. de Jong
Other members:	Prof.dr. P.J.F. Groenen
	Prof.dr. M.-F. Moens
	Dr. D.E. Losada
Copromotor:	Dr.ir. F. Frasincar

Erasmus Research Institute of Management - ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam
Internet: http://www.erim.eur.nl

ERIM Electronic Series Portal: http://repub.eur.nl/pub

ERIM PhD Series in Research in Management, 369
ERIM reference number: EPS-2015-369-LIS
ISBN: 978-90-5892-423-0

Design: B&T Ontwerp en advies www.b-en-t.nl

This publication (cover and interior) is printed by haveka.nl on recycled paper, Revive®.
The ink used is produced from renewable resources and alcohol free fountain solution.
Certifications for the paper and the printing production process: Recycle, EU Flower, FSC, ISO14001.
More info: http://www.haveka.nl/greening

SIKS Dissertation Series No. 2015-26
The research reported in this thesis has been carried out in cooperation with SIKS, the Dutch Research School for Information and Knowledge Systems (http://www.siks.nl)

Preface

> "*A journey is best measured in friends, rather than miles.*"
> Tim Cahill (1944)

Journeys are all about getting to destinations that may at first seem rather distant and vague. In my experience, this holds true for a Ph.D. trajectory as well. A Ph.D. trajectory is a long journey, with the successful defense of a thesis being its intended destination. This destination may at first be nothing but a vague dot on the horizon – the outline and contents of a thesis, as well as the conditions for its defense, are typically shaped on the go. Experimental results, feedback from others, and new insights tend to lead to new research questions that demand further investigation before a thesis can be considered complete, or at the very least defendable. My own Ph.D. trajectory, albeit my personal journey, has not been a solitary one. Many people have helped shape the destination of my Ph.D. trajectory, and for that I am ever grateful.

In the first place, I would like to thank my promotors Prof.dr.ir. Uzay Kaymak and Prof.dr. Franciska de Jong, as well as my copromotor Dr.ir. Flavius Frasincar for guiding me through my Ph.D. trajectory. Uzay, thank you for allowing me to follow my own path, for challenging me not to focus on low-hanging fruit alone, and for stimulating me to set up my own line of research and to develop a coherent thesis in the process. Franciska, you have never ceased to amaze me with your spot-on feedback whenever needed, in spite of your own hectic schedule. Thank you for giving those ever subtle and often crucial nudges in the right direction, even while you were cooking dinner – yes, I did hear the pots and pans in the background when you gave your comments on one of our papers over the phone. Flavius, I know that I must have driven you crazy at first because of my just-in-time handling of deadlines, especially combined with my reluctance of showing any draft at all before being completely satisfied with it. I am thankful for your endless patience, and for how you have helped me make my Ph.D. trajectory a productive one nonetheless. Thank you for being the amazingly friendly, stimulating, and resourceful daily supervisor that you have been to me.

Furthermore, I would like to express my gratitude towards the members of my doctoral committee. Thank you for taking the time for critically reviewing my thesis in light of your own disparate areas of expertise that this dissertation touches upon. I am honored to have you in my committee.

This dissertation would not have existed in its current form without the involvement of the larger scientific community. I am thankful for the numerous anonymous reviewers that have provided valuable feedback on my manuscripts, as this feedback has helped further improve the manuscripts that now constitute this dissertation. Furthermore, I have had the privilege of meeting inspiring colleagues from all over the world at various conferences. Their presentations, their feedback, and our informal conversations have often sparked new ideas. One of such ideas has led to a fruitful collaboration with Dr. David Losada and Jose Manuel Gonzalez Chenlo from the University of Santiago de Compostela. David and Jose Manuel, I am happy that our paths crossed at the CIKM 2011 conference in Glasgow. It has been a great pleasure collaborating with you.

Besides the involvement of the larger scientific community, another crucial enabler for this dissertation has been the support of various organizations. I am thankful for the substantial funding and academic ecosystems provided by the Erasmus Studio, the Econometric Institute, the Erasmus Research Institute for Management (ERIM), the Dutch Research School for Information and Knowledge Systems (SIKS), and the Infiniti project on Information Retrieval for Information Services in the Dutch national program COMMIT.

Concrete and practical support for the day-to-day work that came with my Ph.D. trajectory has been provided by the indispensable supporting staff of the Econometric Institute. Antonia, thank you for always making sure that the tenth floor was a clean and pleasant working environment, in spite of my tendencies of making a mess of both my desk and my chair. Marjon and Anneke, thank you for helping me out with the craziest things, ranging from taking care of several dozens of teaching assistant contracts to helping me fill in the simplest forms that, time and time again, seemed to hold so many mysteries for me. Marianne, Ursula, Elli, and Carien, thank you for making my life at university so much easier due to all of your efforts at the secretariat and in the office management.

Keeping up one's spirits can be a challenge every now and then when undertaking a long journey like a Ph.D. trajectory. In this respect, I have been very lucky to be able to share my journey with a bunch of amazing fellow Ph.D. candidates. First, being able to share most of my journey with my twin brother Frederik has been priceless. Frederik, I could not have wished for a better office mate. It has not only been great fun to share an office with you at the tenth floor – it has been rather convenient as well. We have surely been the perfect occasional substitutes for one another at conferences and lectures...

Even though our office was officially a two-man office, it seemed to double as secondary office of two great colleagues and friends. Damir, your sense of humor, your contagious enthusiasm for new concepts, and your inseparable energy drinks have always helped create a unique atmosphere in the office – relaxed, yet strangely productive. I am happy to have both you and Frederik as my paranymphs. Wim, even the most mundane parts of a Ph.D. trajectory seemed to turn into something hilarious with you around. Thank you for, amidst all hilarity, always challenging my findings and beliefs with your witty and inquisitive nature, and for always putting things into the right perspective. Many more colleagues have made my time at Erasmus University Rotterdam enjoyable. In particular, I would like to thank Rui, Viorel, Milan, Kim, Charlie, Nalan, Tommi, and Yingqian for all the good times we have had at university as well as at conferences.

University is of course much broader than one's own department and direct colleagues. The interdisciplinary lunch meetings of the Erasmus Studio have helped broaden my perspective beyond my own department and field of expertise and have as such been an enriching experience. Furthermore, thanks to my fellow board members, I have had an amazing time in the board of the Erasmus Ph.D. Association Rotterdam (EPAR).

Special thanks go to the students of the Economics & Informatics, Econometrics, and International Business Administration programmes. Educating you in information technology has been such a rewarding experience. In particular, I would like to thank the students with whom I have worked more intensively in the past few years. Bas, Frank, Paul, Daniella, Malissa, Gino, Milan, Maarten, Ferry, and Ewout, supervising you in the context of your seminars and theses has been a great experience, and your hard and devoted work has been truly inspiring. I am proud of how we have established a fruitful collaboration in the process, resulting in various scientific publications.

Support from my friends and family has been indispensable to me during my Ph.D. trajectory. I am thankful to my parents for dealing with my flexible definitions of day and night, for picking me up from my desk every now and then, for putting up with my long shower sessions that often helped clear my mind, for providing me with plenty of food for thought, and for always having a sunny spot available for me to read and review in. Last, to all of you close and dear to me, at times, it may have seemed as though everything revolved around my pursuit of a Ph.D. degree. Thank you for reminding me that there is so much more to life than obtaining a Ph.D. degree. Thank you for sticking with me during my journey. This journey has now come to an end, and new journeys await.

Rotterdam, July 2015
Alexander Hogenboom

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 7

List of Tables

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

List of Algorithms

Chapter 4

Chapter 5

Chapter 1

Introduction

Challenging economic conditions, speculative bubbles for cryptocurrencies like Bitcoin, and electronic word-of-mouth phenomena on social media have recently demonstrated how today's markets are affected by people's sentiment, i.e., by people's moods or opinions. We have now reached a point where sentiment plays a pivotal role in various business and economic processes – consumer confidence is strongly related to the realization of economic conditions (Howrey, 2001; Ludvigson, 2004; Vuchelen, 2004), Bitcoin exchange rates are extremely sensitive to the craze of the day (Kristoufek, 2013), and sales (Rui et al., 2013; Yu et al., 2012), stock ratings (Yu et al., 2013), and reputation (Jansen et al., 2009) are influenced by subjective tweets, reviews, and other social media content. As such, keeping track of their stakeholders' sentiment is crucial for today's decision makers.

A traditional way of keeping track of the sentiment of one's stakeholders is to perform a representative survey that focuses on opinions on the current situation and expectations for the (near) future. Such a procedure is typically applied in order to compile macroeconomic indicators that capture economic sentiment, such as the University of Michigan Consumer Sentiment Index (CSI) or the Consumer Confidence Index (CCI) (Ludvigson, 2004). However, people's expectation formation is thwarted by structural, psychologically driven distortions (Bovi, 2009) – when respondents consider survey questions to be vague or hard to assess, they tend to provide biased answers (Tversky and Kahneman, 1974). Furthermore survey panels tend to be rather small and encompass different respondent samples over time. This complicates the generalizability of survey findings, as observed sentiment shifts may be largely driven by differences in respondent samples (van Oest and Franses, 2008). Moreover, survey-based methods for tracking sentiment cannot tap into *many* people's sentiment *right now.* Yet, in today's complex and dynamic markets, it is crucial for decision makers to understand and react to ever-changing circumstances in a timely and effective manner (Hogenboom et al., 2015d).

Fortunately, we live in an era in which digital traces of sentiment are ubiquitous. The Web as we know it today is a network of more than 555 million Web sites, with over two billion users (Pingdom, 2012). Every second, these users collectively enter about 50,000 search terms in query engines like Google (Internet Live Stats, 2014) and thus provide indirect proxies for their sentiment (Della Penna and Huang, 2009; Kristoufek, 2013; Preis et al., 2010; Vosen and Schmidt, 2011). Moreover, today's Web users' sentiment is revealed in a more explicit and specific manner in the thousands of reviews, blog posts, and tweets that are posted every second of the day. Besides the reviews, one in three blog posts (Melville et al., 2009) and one in five tweets (Jansen et al., 2009) discuss products or brands, and hence contain valuable information. However, in this era of Big Data, the abundant and ubiquitous user-generated content is often unstructured, scattered across the Web, and ever-expanding, thus rendering information extraction by manual analysis of all available data unfeasible (Madden, 2012). Tools for automated sentiment analysis of textual data can provide timely and effective support for decision making processes.

1.1 Automated Sentiment Analysis

Systems that perform automated sentiment analysis are mainly concerned with the extraction of subjective information from natural language text. This task poses three main challenges, i.e., the identification of relevant pieces of text (Mei et al., 2007; O'Hare et al., 2009; Zhang and Ye, 2008), the analysis of relevant textual data for conveyed sentiment (Cambria et al., 2013; Feldman, 2013; Liu, 2012; Pang and Lee, 2008), and the analysis of potential effects of the identified sentiment on entities of interest (e.g., products or brands) (Amigo et al., 2013; Jansen et al., 2009). This dissertation is specifically focused on the task of analyzing the sentiment conveyed by (presumably) appropriate textual data.

The analysis of sentiment conveyed by pieces of natural language text can serve various goals. Some sentiment analysis methods focus on distinguishing subjective text segments from objective ones (Wiebe et al., 2004) or on summarizing opinions (Lerman et al., 2009; Titov and McDonald, 2008), but the most common goal of sentiment analysis approaches is the identification of the polarity of words, sentences, text segments, or documents (Liu, 2012; Pang and Lee, 2008). This dissertation deals with the latter task of analyzing the polarity of natural language text. This task is commonly treated as a binary classification problem, which involves classifying the polarity of text as either positive or negative. However, more polarity classes – such as classes of neutral or mixed polarity, or star ratings ranging from, e.g., one to five stars – may be considered. The degree of positivity or negativity of a piece of text is another potential outcome of a polarity analysis.

Existing polarity analysis techniques stem from the fields of natural language processing, computational linguistics, and text mining, and range from machine learning methods to rule-based methods. Machine learning methods involve training of models on specific collections of documents (i.e., corpora) by means of mostly supervised methods that exploit patterns in vector representations of natural language text (Taboada et al., 2011). The most common and most useful features in these vectors indicate the presence or frequencies of specific words or word groups and constitute a so-called *bag-of-words* representation of text (Pang et al., 2002; Pang and Lee, 2004). These word-based features are often enriched with part-of-speech (POS) information, thus enabling the distinction between (types of) nouns, verbs, adjectives, and adverbs (Liu, 2012). Machine learning polarity analysis methods tend to perform comparably well on the corpora for which they have been optimized (Chaovalit and Zhou, 2005; Kennedy and Inkpen, 2006; Liu, 2012; Pang and Lee, 2008; Taboada et al., 2011), but they require a lot of (annotated) training data as well as training time in order to attain this competitive level of performance.

Compared to machine learning methods, rule-based polarity analysis approaches tend to be rather robust across domains and texts in terms of accuracy (Taboada et al., 2011). Rule-based methods mostly rely on lexicons that list words and their associated sentiment scores. The sentiment scores of words in a text are typically combined (e.g., summed or averaged) in accordance with predefined rules and assumptions in order to obtain a text's overall sentiment score, which can be used as a proxy for the text's polarity. This process renders the motivation for assigning a particular polarity to a text rather transparent, as opposed to black-box machine learning approaches. In the sentiment scoring process, negation (Heerschop et al., 2011c,d; Hogenboom et al., 2011a,b) or intensification (Taboada et al., 2011) of sentiment conveyed by words may be accounted for. Syntactic patterns based on POS information may be exploited as well, in order to focus the analysis on specific parts of a text that may be opinionated, such as adjectives that are preceded by adverbs (e.g., "*very good*") (Turney, 2002). As such, rule-based approaches allow for intuitive ways of incorporating linguistic analysis into the polarity analysis process.

1.2 Levels of Linguistic Analysis

Natural language text can be analyzed at various levels (Jurafsky and Martin, 2000; Liddy, 2003), as visualized in Figure 1.1. The distinction between levels may not be clear-cut per se as it is the combination of these levels that helps convey meaning (Liddy, 2003). Yet, even though humans use all levels in order to gain understanding, this is not necessarily the case for systems that automatically process natural language (Liddy, 2003).

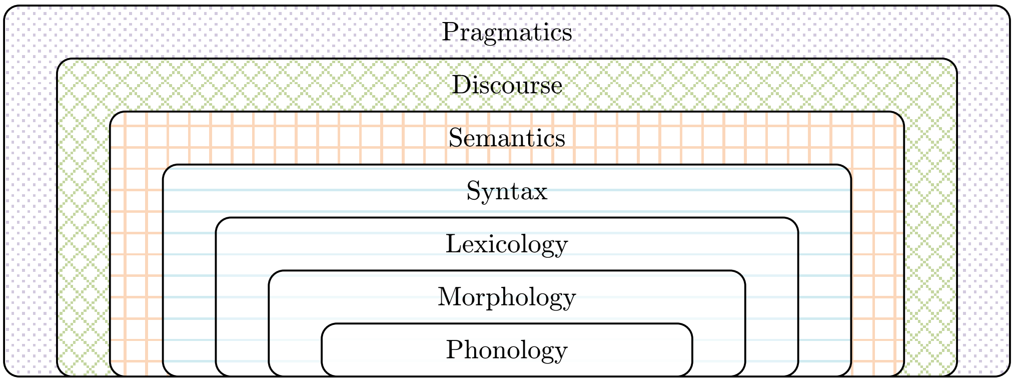


Figure 1.1: Levels of linguistic analysis.

The lowest level of linguistic analysis deals with phonetics and – especially – phonology. Phonetics comprises the study of the specific sounds that can be physically produced by humans. Phonology is concerned with how (a subset of) these sounds can be used in a language in order to encode meaning, thus enabling for instance the subtle difference between the singular "*man*" and the plural "*men*".

A higher level of analysis is concerned with morphology, which deals with how morphemes, i.e., the smallest meaning-carrying units of a language, constitute words. In addition to a root, a word may contain one or more affixes. Each of these morphemes contributes to conveying a word's meaning. For example, the word "*preselected*" consists of the root "*select*", a prefix "*pre*" that signals that the selection has been made in advance, and a suffix "*ed*" that signals the root verb's perfect tense. Morphology builds upon phonology. For instance, even though the plural of "*dog*" has a suffix "*s*" (i.e., "*dogs*"), the phonological constraints for the English language dictate that the plural of "*match*" should have a suffix "*es*" rather than "*s*", because the "*tchs*" sound of "*matchs*" is invalid.

The subsequent level of lexical analysis deals with interpreting the meaning of individual words. In this process, a word can be assigned a lexical category, denoting its POS. The POS of a word plays an important role in conveying the word's meaning. For instance, the adverb "*well*" implies success, whereas the noun "*well*" refers to a deep hole that has been dug in the ground in order to provide for, e.g., water.

The syntactic level of linguistic analysis is concerned with the meaning conveyed by combined words that form phrases or sentences. Syntactic dependency relationships between words in a text can bring about differences in meaning when including or leaving out particular words, or when changing the word order. For instance, in the phrase "*I do not like the awful simplicity of the plot*", "*not*" negates "*like*", thus conveying that something is disliked, and "*awful*" amplifies the negative connotations of "*simplicity*".

Leaving out the words "*not*" and "*awful*" would completely change the meaning of the phrase. Furthermore, the phrases "*The dog bites a man*" and "*The man bites a dog*" convey the same information on a lexical, morphological, or phonological level, whereas their differences in word order result in completely different meanings on a syntactic level.

When analyzing text on a semantic level, the interactions of word-level meanings are dealt with. Some words have multiple senses, and their intended sense depends on the meaning of the context in which these words occur. Semantic analysis deals with such word sense disambiguation tasks. For example, the adjective "*cool*" could refer to a fairly low temperature, if the context in which it occurs discusses meteorological conditions, whereas it may express fashionable attractiveness or impressiveness in a review. Similarly, a "*bank*" can represent two disparate concepts when collocated with "*water*" and "*money*".

A higher level of linguistic analysis is concerned with discourse analysis. Discourse analysis deals with how meaning is built up in the larger communicative process. The premise is that each phrase, sentence, paragraph, or other discourse unit has a specific role in conveying the overall message of a text. Identifying a text's rhetorical structure and the roles that text fragments have within this structure can provide valuable additional information, as it can help put the meaning of text fragments in the right perspective. Background or off-topic information may for instance be less relevant than or tangential to the main ideas and conclusions presented in a text.

The highest level of linguistic analysis deals with pragmatics, i.e., the purposeful use of language in specific situations, such that the meaning of a piece of text per se may differ from its meaning in practical use. The focus here is on enriching text with meaning that is not actually encoded in the text itself, but rather requires real-world knowledge. For example, the phrase "*This movie is as entertaining as Wild Wild West*" can only be understood to its full extent when incorporating the real-world knowledge of "*Wild Wild West*" being a movie from the late 1990's that received Golden Raspberry Awards for, among others, Worst Picture and Worst Screenplay.

1.3 Research Objectives

Most existing approaches to automated sentiment analysis focus on the lower levels of linguistic analysis by making use of mostly morphological, lexical, and syntactic information (Liu, 2012; Pang and Lee, 2008). However, as the utilization of more, higher levels of linguistic analysis can improve a system's understanding of text (Liddy, 2003), the analysis of the polarity of text may be more accurate if it additionally accounts for semantics, as well as for the rhetorical structure of text as identified by means of discourse analysis.

Pragmatics could be exploited as well, yet doing so would require the incorporation of real-world knowledge from beyond the natural boundaries of a text. Therefore, in order to better utilize the potential of the information contained *within* text in the analysis of its polarity, the problem statement underlying this dissertation is:

How can semantic and structural aspects of text complement low-level linguistic information in the analysis of the polarity of text?

The low-level linguistic information referred to in this problem statement includes the typical morphological and lexical cues for a text's polarity, as well as their associated POS information and syntactic dependencies that capture, e.g., negation or amplification. Such morphological and lexical cues can be, for instance, (roots of) words or sequences of characters representing emotions – i.e., emoticons.

Semantics can be involved in the polarity analysis of text by accounting for the interactions between cues for sentiment on a semantic level, thus capturing how the sentiment conveyed by some cues depends on the meaning of other cues. For example, the sentiment conveyed by words may depend on co-occurring emoticons. Another potentially fruitful way in which semantics can complement low-level linguistic information is the exploitation of semantic relations like synonymy and antonymy in order to propagate sentiment-related information from known cues to other, semantically related cues with an – a priori – unknown sentiment. This could be particularly useful in a multi-lingual setting in which sentiment-related information is scarce for (some of) the targeted languages.

In addition to accounting for semantic aspects of text, discourse analysis can be applied in order to be able to guide the polarity analysis process by a text's rhetorical structure. This would allow for a distinction between the sentiment conveyed by, e.g., conclusions and the sentiment conveyed by, e.g., background information. However, scalability issues may need to be addressed in order to accomplish this, since automated discourse analysis is a computationally intensive process.

Mechanisms for accounting for semantic interactions, as well as for guiding the polarity analysis process by the rhetorical structure of text can be explicitly modeled by means of sets of rules. This allows for intuitive ways of incorporating such mechanisms in rule-based polarity analysis systems. However, semantic and structural aspects of text can also be captured by features that can prove beneficial for machine learning methods.

Various facets of the problem statement underlying this dissertation are addressed in the scientific articles (either published or accepted for publication) that constitute this dissertation. These articles deal with six distinct research questions.

Question 1: To what extent can polarity classes be identified by means of sentiment scores that stem from low-level linguistic analysis?
As most existing work on polarity classification focuses on the lower levels of linguistic analysis and as such makes use of morphological, lexical, and syntactic information (Liu, 2012; Pang and Lee, 2008), an initial assessment of these levels' merits for polarity classification is an intuitive first step towards assessing the added value of higher levels of linguistic analysis. Therefore, in this dissertation, such an assessment is performed through an evaluation of how the sentiment conveyed by individual words and their modifiers relates to polarity classes that capture a universal measure of the sentiment that authors of texts have intended to convey, i.e., author-provided star ratings that range from one to five stars. Models of varying complexity are considered in order to capture this relation, i.e., a monotonically increasing step function, a naive Bayes method, and a support vector machine. This analysis is done for English movie reviews as well as for Dutch movie reviews, in order to assess whether the findings are language-dependent.

Question 2: How do emoticons interact with sentiment-carrying content on a semantic level and how can this be exploited in polarity classification?
Over the years, people have embraced the usage of so-called emoticons in user-generated content, as a means of providing – virtual – visual cues on how words in their written text should be interpreted. The meaning of a piece of text may as such depend on the presence and implications of emoticons, which thus contain potentially valuable polarity-related information. This advocates the need to study how emoticons interact on a semantic level with (the sentiment conveyed by) the words in a piece of written text, and to subsequently capture the mechanics of this interaction in a rule-based polarity classifier. In the work discussed in this dissertation, such a classifier is evaluated on various collections of documents, i.e., on Dutch tweets and forum messages, as well as on English app reviews. This evaluation is focused on assessing to what extent polarity classification performance can be improved by accounting for the modeled semantic interactions of emoticons with textual content.

Question 3: In what way can semantics assist in polarity classification in a multi-lingual setting?
The ever-growing amount of data in different languages renders multi-lingual polarity classification increasingly important. An intuitive approach would involve analyzing the polarity of texts that have been machine-translated to a reference language for which a polarity classifier is available. Yet, semantics may be lost in translation (Mihalcea et al., 2007) and the original content may be inaccurately represented in the reference language.

Devising a new polarity classifier for another language is an attractive alternative approach, provided that tools for performing linguistic analysis on morphological, lexical, and syntactic level are available for that language. Whereas machine learning methods would require large amounts of annotated training data for a new language, rule-based approaches that rely on sentiment lexicons can be generalized from a reference language to another language relatively easily by using another lexicon. Such a lexicon can be constructed for a new language by carefully considering semantic relations between and within languages. This dissertation deals with various ways of exploiting such relations in order to create a polarity classifier for Dutch, given a polarity classifier for English.

Question 4: How can rule-based polarity classification be guided by the rhetorical structure of text?

A text's rhetorical structure can provide detailed information on the role that specific text segments play in conveying the meaning of the text. Such information can help making rule-based distinctions between important text segments and less important ones in terms of their contribution to the sentiment conveyed by a piece of text as a whole. In work covered by this dissertation, the importance of text segments is identified based on the rhetorical structure of text as automatically identified by applying the Rhetorical Structure Theory (RST) of Mann and Thompson (1988) to various units of analysis and with various levels of granularity. The aim here is to extract information from a text's rhetorical structure and to subsequently combine this information with sentiment-related information as extracted by means of lower levels of linguistic analysis. This approach is formalized in a rule-based classifier, that moreover accounts for semantics by disambiguating word senses and retrieving sentiment scores for the identified senses from a semantically enabled sentiment lexicon. This classifier is applied to a corpus of English movie reviews in order to assess the performance of polarity classification guided by the rhetorical structure of text.

Question 5: In what way can the rhetorical structure of text be exploited in large-scale polarity analysis?

The deep linguistic analysis needed in order to automatically identify the rhetorical structure of a text is computationally intensive. This may thwart the applicability of structure-guided polarity analysis in large-scale scenarios of use. Therefore, this dissertation covers work that investigates how a focused, rule-based, structure-guided analysis of the sentiment conveyed by selected text segments can improve the performance of a polarity analysis tool in a large-scale polarity ranking task on a collection of English blog posts.

Question 6: How can the rhetorical structure of text be taken into account by a machine learning approach to polarity classification?
The rhetorical structure of a piece of natural language text can be accounted for in an intuitive way by a rule-based approach to polarity classification, because such an approach allows for the polarity analysis to differentiate between text segments based on their respective rhetorical roles. However, machine learning methods can achieve a more competitive polarity classification performance than rule-based methods (Chaovalit and Zhou, 2005; Kennedy and Inkpen, 2006; Liu, 2012; Pang and Lee, 2008; Taboada et al., 2011). Machine learning approaches to sentiment analysis typically use features that stem from morphological, lexical, or syntactic linguistic analysis, with the most common and most valuable features representing the (frequencies of) occurrence of specific words or word groups (Liu, 2012; Pang and Lee, 2008). In work included in this dissertation, features that capture a text's rhetorical structure are proposed and evaluated for their usefulness in a machine learning approach to polarity classification of collections of English reviews in various domains. Semantics are accounted for as well, by applying word sense disambiguation and subsequently representing the words that occur in the reviews as the semantic concepts to which their senses belong.

1.4 Contributions

The contributions of this dissertation are manifold. First, commonly applied morphological, lexical, and syntactic processing is utilized in order to analyze the sentiment conveyed by individual words and their modifiers (Bal et al., 2011; Hogenboom et al., 2012a, 2014a). The contribution of these endeavors is not in the application of these common natural language processing techniques per se, but rather in the proposal and evaluation of various methods of mapping the resulting sentiment scores to polarity classes that capture a universal measure of authors' intended sentiment. The most advanced method proposed in this work incorporates a representation of the sentiment-carrying content (Hogenboom et al., 2012b) into the polarity classification process. Another contribution of this work lies in the analysis of the extent to which the sentiment-related information obtained through morphological, lexical, and syntactic processing of text can separate universal classes of intended sentiment for movie reviews that have been written in two different languages, i.e., Dutch and English.

Other work reported on in this dissertation explores how emoticons interact on a semantic level with the (sentiment conveyed by) words in a text, and how this can be exploited when classifying the polarity of pieces of text (Hogenboom et al., 2013a, 2015a).

The key contribution of this particular work lies in the analysis of the role that emoticons play in conveying a text's overall sentiment, as well as in the proposal and evaluation of a novel, rule-based method that exploits emoticons in lexicon-based polarity classification of documents from various domains, in both Dutch and English. Rather than accounting for emoticons on a lexical level (Thelwall et al., 2010), this work shows how to jointly use explicit textual cues and emoticons when classifying the polarity of a text, by incorporating linguistic analysis on a semantic level into the polarity classification process.

The added value of accounting for semantics is further elaborated on in other work covered by this dissertation. This work explores how a sentiment lexicon can be constructed for a new language by carefully considering semantic relations between and within languages (Hogenboom et al., 2014b). The main contribution of this work lies in a novel sentiment mapping method that exploits semantic relations between language-specific semantic lexicons in order to construct a sentiment lexicon for a target language, i.e., Dutch, by making use of an existing sentiment lexicon for a reference language, i.e., English. The effectiveness of this method is compared with the effectiveness of a method that focuses on semantic relations within, rather than across languages (Heerschop et al., 2011b), as well as with the effectiveness of a machine-translation approach that does not specifically account for semantics. These research efforts provide insight into the importance of semantics for polarity classification in a multi-lingual setting.

In addition to the work discussed above, this dissertation covers work in which not only semantic analysis, but also discourse analysis is applied in order to improve the polarity classification performance of methods that account for morphological, lexical, syntactic, and semantic information only (Heerschop et al., 2011a; Hogenboom et al., 2010b, 2015b). The main contributions of this work include the proposal and evaluation of methods of accounting for a text's rhetorical structure – as identified by an automated RST-based analysis – on various levels of detail and for various units of analysis. Another contribution lies in the proposed weighting schemes that enable a rule-based polarity classifier to make a fine-grained distinction between text segments and their conveyed sentiment, based on these segments' identified rhetorical roles. An analysis of the performance of this classifier on a set of English movie reviews provides insight into how a text's rhetorical structure can best be accounted for. Moreover, in other work discussed in this dissertation, the scalability of RST-guided rule-based polarity analysis is improved by guiding the polarity analysis process by a shallow analysis of the rhetorical structure of only a small fraction of all text (Chenlo et al., 2013, 2014). An evaluation of the performance of the resulting polarity analysis tool in a large-scale polarity ranking task for English blog posts provides insight into the feasibility of large-scale RST-guided polarity analysis.

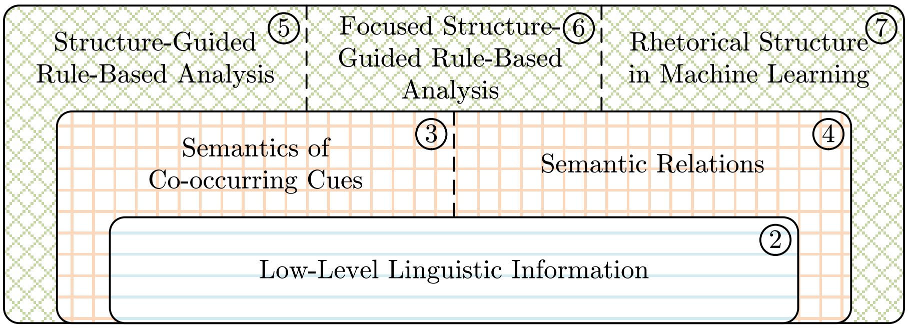

Structure-Guided Rule-Based Analysis
5
Focused Structure-Guided Rule-Based Analysis
6
Rhetorical Structure in Machine Learning
7
Semantics of Co-occurring Cues
3
Semantic Relations
4
Low-Level Linguistic Information
2

Figure 1.2: Content covered by the main chapters of this dissertation. The circled numbers represent chapter numbers. The markings correspond with those in Figure 1.1.

A final contribution of this dissertation lies in the proposed structure-based features that facilitate a richer vector representation of natural language text, which should contribute to a more accurate classification of its polarity by means of a machine learning classifier (Hogenboom et al., 2015c). These features capture the extent to which text conveys sentiment as well, as this has been shown to be an important cue in sentiment analysis (Mangnoesing et al., 2012). The evaluation of a machine learning polarity classifier that uses the proposed features enables insight into the importance of accounting for structural aspects of content in a – performance-wise comparably competitive – machine learning approach to polarity classification, such that not only rule-based, but also machine learning systems for automated sentiment analysis can be used more effectively for supporting decision making processes.

1.5 Outline

The next six chapters of this dissertation are based on scientific articles (either published or accepted for publication) that address the six research questions posed in Section 1.3. Figure 1.2 visualizes how these chapters build upon one another by exploring various ways of deploying increasingly more levels of linguistic analysis in the polarity analysis of text.

Chapter 2 addresses the first research question by assessing the extent to which sentiment-related information obtained through morphological, lexical, and syntactic processing can be used for classifying the polarity of reviews. As such, Chapter 2 evaluates the usefulness of commonly used information obtained through low-level linguistic analysis.

Subsequently, Chapters 3 and 4 investigate the merits of additionally performing a semantic analysis when determining the polarity of various types of documents. Chapters 3 and 4 thus deal with the second and third research question, respectively, by including an extra level of linguistic analysis in the polarity classification process.

The merits of additionally incorporating discourse analysis into the polarity analysis process are investigated in Chapters 5, 6, and 7, in which not only semantic, but also structural aspects of content are accounted for in the analysis of the polarity of natural language text. In Chapter 5, a rule-based structure-guided polarity classification approach is proposed and evaluated, thus addressing the fourth research question. In order to answer the fifth research question, the developed rule-based structure-guided polarity analysis approach is adapted in Chapter 6 in order for it to be more suitable for large-scale polarity analysis. Chapter 7 deals with the sixth research question by proposing features that capture semantic and – above all – structural aspects of content, and by subsequently evaluating their usefulness in a machine learning approach to polarity classification.

Chapter 8 summarizes the main findings of the research covered by this dissertation and additionally presents concluding remarks regarding the underlying problem statement. An outlook on the implications of these findings and envisaged directions for future research then conclude this dissertation.

Chapter 2

Mapping Word-Based Sentiment Scores to Intended Sentiment[*]

WITH *consumers generating increasingly more content describing their experiences with, e.g., products and brands in various languages, information systems that can monitor a universal, language-independent measure of people's intended sentiment are crucial for today's businesses. In order to facilitate sentiment analysis of user-generated content, we propose to map the sentiment conveyed by the words used in unstructured natural language text to universal star ratings that capture the intended sentiment. For these mappings, we consider a monotonically increasing step function, a naive Bayes method, and a support vector machine. We demonstrate that the way in which sentiment-carrying words reveal intended sentiment differs across our collections of Dutch and English texts. Moreover, our experimental results indicate that language-specific sentiment scores based on sentiment-carrying words can separate universal classes of intended sentiment from one another only to a limited extent – semantic and structural aspects of content appear to play an important role in conveying an author's intended sentiment.*

[*]This chapter is based on the article "A. Hogenboom, M. Bal, F. Frasincar, D. Bal, U. Kaymak, and F. de Jong. Lexicon-Based Sentiment Analysis by Mapping Conveyed Sentiment to Intended Sentiment. *International Journal of Web Engineering and Technology*, 9(2):125–147, 2014."

2.1 Introduction

Today's consumers are increasingly more inclined to share their opinions or experiences with, e.g., products and brands through the Web in the language of their preference. By now, one in three blog posts (Melville et al., 2009) and one in five tweets (Jansen et al., 2009) discuss products or brands. As anyone can nowadays write reviews and blogs, post messages on discussion forums, or publish whatever crosses one's mind on Twitter at any time, today's businesses face a continuous flow of an overwhelming amount of multi-lingual data of all sorts, containing traces of valuable information – consumers' sentiment with respect to products, brands, and so on. In this wealth of user-generated content, explicit information on user opinions is often hard to find, confusing, or overwhelming (Pang and Lee, 2008). As such, the abundance of sentiment-carrying user-generated content renders automated information monitoring tools for sentiment crucial for today's businesses.

Such information monitoring tools rely on sentiment analysis techniques, stemming from natural language processing, computational linguistics, and text mining. The goal of most sentiment analysis approaches is to determine the polarity of natural language text. Typical methods involve scanning a text for cues – e.g., words – signalling its polarity. Most state-of-the-art methods are machine learning approaches. Nevertheless, the use of sentiment lexicons, i.e., lists of words and their associated sentiment, possibly differentiated by Part-of-Speech (POS) and/or meaning (Baccianella et al., 2010), has gained attention in recent work (Cesarano et al., 2006; Devitt and Ahmad, 2007; Ding et al., 2008; Heerschop et al., 2011a,c; Hogenboom et al., 2012b; Taboada et al., 2011), as lexicon-based approaches have been shown to have a more robust performance across domains and texts than machine learning methods (Taboada et al., 2008). Additionally, lexicon-based methods allow for intuitive ways of accounting for other cues for sentiment – e.g., emoticons (Hogenboom et al., 2013a) – as well as for incorporating deep linguistic analysis into the sentiment analysis process, for instance by accounting for structural or semantic aspects of text (Chenlo et al., 2013; Heerschop et al., 2011a).

Existing sentiment analysis methods typically consist of language-specific components such as sentiment lexicons, or components for, e.g., identifying the lemma or POS of words. Each language-specific sentiment analysis approach typically produces sentiment scores for texts in its reference language, ranging from, e.g., -1 (negative) to 1 (positive). Intuitively, such scores should be meaningful and comparable across languages. Therefore, many existing methods of analyzing sentiment in a multi-lingual setting make use of language-specific sentiment analysis approaches, and subsequently treat all language-specific sentiment scores equally in a cross-lingual analysis of sentiment (Bal et al., 2011).

However, sentiment scores have been shown not to be directly comparable across languages, as they tend to be affected by several language-specific phenomena, such as expressions or culture-dependent semantics (Bal et al., 2011; Wierzbicka, 1995a,b).

The language-specific sentiment scores produced by existing sentiment analysis methods typically reflect the sentiment conveyed by the cues in the natural language content, which is not necessarily the sentiment which authors of such content have intended to convey. Therefore, we propose to map language-specific sentiment scores to a universal, language-independent measure of people's intended sentiment, i.e., star ratings. The number of stars assigned to a text typically reflects the extent to which the author (e.g., a reviewer) intends to convey positive sentiment with respect to the subject of the text (e.g., a reviewed product). As universal star ratings capture people's intended sentiment rather than the language-dependent sentiment conveyed by natural language text, these star ratings can be used as culture-free analytical tools for analyzing people's sentiment.

Star ratings are, however, not always available. For instance, opinionated blog posts or tweets are not typically assigned scores by their respective authors in order to signal their intended sentiment. In this light, a major challenge is to automatically determine the star rating associated with reviews based on cues in the actual natural language content. In this chapter, we aim to gain insight in the relation between language-specific scores of sentiment conveyed by the words in natural language content on the one hand, and universal star ratings of intended sentiment on the other hand. As such, we aim to benefit from the robust and fine-grained type of analysis that traditional, lexicon-based sentiment analysis techniques offer (Taboada et al., 2008), while using universal star ratings to capture people's intended sentiment.

The remainder of this chapter is structured as follows. First, we discuss related work on multi-lingual sentiment analysis in Section 2.2. Then, in Section 2.3, we propose several methods for mapping language-specific word-based sentiment scores to universal classifications of intended sentiment in order to facilitate more meaningful analyses of people's true sentiment. An evaluation of our methods on Dutch and English documents is presented in Section 2.4. Last, we conclude in Section 2.5.

2.2 Analyzing Sentiment in a Multi-lingual Setting

In an extensive literature survey on sentiment analysis (Pang and Lee, 2008), the current surge of research interest in systems dealing with opinions and sentiment is attributed to the fact that, despite today's users' hunger for and reliance upon on-line recommendations, explicit information on user opinions is hard to find, confusing, or overwhelming.

Many sentiment analysis approaches exist (Cambria et al., 2013; Feldman, 2013; Liu, 2012; Pang and Lee, 2008), yet the relation between language-specific sentiment scores stemming from such approaches and the actually intended sentiment has been relatively unexplored.

Some existing sentiment analysis approaches exploit (generic) lists of words and their associated sentiment, i.e., sentiment lexicons, when determining the subjectivity or polarity of natural language text, possibly while accounting for negation (Heerschop et al., 2011c; Hogenboom et al., 2011a) or intensification (Taboada et al., 2011) of the sentiment conveyed by specific words. Other methods rely on machine learning techniques in order to exploit patterns in vector representations of text. These vectors are typically bag-of-words representations, in which features may represent the presence (Pang and Lee, 2004) or frequency (Pang et al., 2002) of specific words, parts of words, or word groups.

Lexicon-based approaches have an attractive advantage over machine learning approaches to sentiment analysis in that they have a robust performance across domains and texts (Taboada et al., 2008). Additionally, lexicon-based approaches enable deep, yet computationally-intensive linguistic analysis to be incorporated into the sentiment analysis process (Heerschop et al., 2011a). Moreover, lexicon-based approaches can be generalized relatively easily to other languages by using dictionaries (Mihalcea et al., 2007). On the other hand, lexicon-based methods tend to sacrifice computational efficiency as they often incorporate deep linguistic analysis into the sentiment detection procedures (Heerschop et al., 2011a), and they are typically outperformed by machine learning methods in terms of classification accuracy in specific domains for which machine learning methods can be trained and optimized (Taboada et al., 2008).

This trade-off has inspired hybrid approaches, which combine the classification accuracy and processing speed benefits of machine learning approaches with the robustness of lexicon-based methods. A promising step into this direction has been made with the introduction of a *bag-of-sentiwords* representation (Hogenboom et al., 2012b), where a text is represented by means of a binary vector representation, with the features representing the presence of sentiment-carrying words, retrieved from a general purpose sentiment lexicon. These sentiment-carrying words are assumed to play a crucial role in conveying sentiment, as opinionated texts significantly differ from non-opinionated texts in terms of occurrences of subjective words (van der Meer et al., 2011). The motivation for a binary representation lies in its superiority over a frequency-based representation (Pang et al., 2002) as well as in an assumption that the sentiment conveyed by a text is not so much in the number of times a single word occurs in a text, but rather in the distinct words with a similar semantic orientation.

Today's sentiment analysis systems must be able to deal with an abundance of multi-lingual sentiment-carrying user-generated content in order to facilitate meaningful analyses of, for example, consumer sentiment with respect to products or brands. Existing sentiment analysis approaches focus on determining language-specific sentiment scores for (collections of) documents in selected languages, mainly by applying sentiment analysis techniques tailored to each specific considered language, as different sentiment analysis approaches are required for distinct languages (Boiy and Moens, 2009).

One way of dealing with documents in multiple languages is proposed by Bautin et al. (2008), who apply machine translation in order to convert all of their considered documents into a reference language, i.e., English. Subsequently, sentiment analysis is performed on the translated documents. By doing so, Bautin et al. (2008) assume that the results of the performed sentiment analyses on both the original text and the translated text are comparable and that the errors made by the machine translation step do not significantly influence the outcome of the sentiment analysis process. However, the quality of the sentiment analysis on the translated documents in fact does depend on the translation quality in terms of, e.g., the accuracy of the representation of the original text at a semantic level.

As machine translation approaches to sentiment analysis clearly have their limitations, existing research on dealing with a multi-lingual setting when analyzing sentiment typically targets the sentiment scoring problem for each considered language separately. Existing work is primarily focused on how to devise sentiment scoring methods for new languages with minimal effort, yet without sacrificing too much accuracy. The focus of existing work varies from creating sentiment lexicons (Hofmann and Jijkoun, 2009; Wan, 2009) to constructing entirely new sentiment scoring frameworks (Abbasi et al., 2008; Boiy and Moens, 2009; Dai et al., 2007a,b; Gliozzo and Strapparava, 2005) for languages other than the reference language.

The resulting scores reflecting the sentiment conveyed by natural language content are not particularly meaningful per se. In recent work, Bal et al. (2011) compare the sentiment conveyed by the natural language content of documents with the sentiment conveyed by their translated counterparts. The experiments of Bal et al. (2011) show that sentiment scores are not directly comparable across languages, as these scores tend to be affected by many different language-specific phenomena. Moreover, other research has shown that there is a cultural dimension to sentiment differences across languages, as every language imposes its own classification upon human emotional experiences, thus rendering sentiment-carrying words in a particular language artifacts of that language rather than culture-free analytical tools (Wierzbicka, 1995a,b).

Therefore, in order for language-specific sentiment scores to be meaningful when analyzing user-generated content for the associated sentiment, we need to map such scores to a universal, language-independent measure of people's intended sentiment. In this chapter, we assume that star ratings reflect people's intended sentiment, as authors of, e.g., reviews can typically quantify their overall verdict by means of such star ratings. In any language, a higher number of stars associated with a text is typically associated with a more positive sentiment of the author towards the topic of this text. As such, star ratings are universal classifications of the sentiment that people actually intend to convey, whereas traditional sentiment scores tend to reflect the sentiment conveyed by the way people express themselves through the words they use.

Intuitively, both measures may be related to some extent, yet to the best of our knowledge, the relation between language-specific word-based sentiment scores and universal sentiment classifications has not been previously investigated. The contribution of our current work lies in investigating how language-specific word-based sentiment scores can be mapped to universal star ratings.

2.3 From Sentiment Scores to Star Ratings

As traditional lexicon-based sentiment analysis techniques are guided by the natural language used in texts, they allow for a fine-grained linguistic analysis of the sentiment conveyed by these texts. In addition, traditional lexicon-based techniques are rather robust as they take into account the actual content of a piece of natural language text, especially when involving structural and semantic aspects of content in the analysis (Chenlo et al., 2013; Heerschop et al., 2011a; Taboada et al., 2011). As such, these lexicon-based sentiment analysis techniques may prove to be useful for analyzing the enormous variety of multi-lingual user-generated content.

Traditional lexicon-based sentiment analysis approaches typically aim to assign sentiment scores to natural language text, ranging from, e.g., -1 (negative) to 1 (positive). In order to support amplification of sentiment, e.g., "very bad" rather than "bad", sentiment scores may also range from, e.g., -1.5 (very negative) to 1.5 (very positive). However, as such scores are not particularly meaningful as they are a quantification of the sentiment conveyed by natural language rather than a language-independent, universal measure of intended sentiment (Bal et al., 2011), a mapping from language-specific scores of conveyed sentiment to universal classifications of intended sentiment is of paramount importance, and a particular contribution of our work.

2.3.1 Sentiment Scoring

As a first step, we propose to compute language-specific sentiment scores by means of a sentiment scoring approach such as the method proposed by Bal et al. (2011). This framework is essentially a pipeline in which each component fulfils a specific task in analyzing the sentiment of a document.

For each supported language, our sentiment analysis framework first prepares documents by cleaning the text – i.e., by converting the text to lowercase, removing diacritics, etcetera. Initial linguistic analysis is subsequently performed by identifying each word's POS as well as by distinguishing sentiment-carrying words and their modifiers (e.g., words that negate or amplify the sentiment conveyed by a word) from words that do not carry any sentiment.

Then, for each sentiment-carrying word t in a document d, the word-level sentiment score ζ_t as well as the strength of its modifier m_t (if any) is retrieved from a sentiment lexicon. Sentiment scores range from -1.0 (negative) to 1.0 (positive), whereas modifiers range from -1.5 (amplified negation) to 1.5 (amplification). If a sentiment-carrying word t is not modified, m_t is set to 1.0. The word-based sentiment score ζ_d of a document d can then be determined by sum-aggregating the (modified) sentiment scores of the individual words and by subsequently normalizing the result for the number of sentiment-carrying words, i.e.,

$$\zeta_d = \frac{\sum_{t \in d} m_t \zeta_t}{|t \in d|}. \tag{2.1}$$

The normalized word-based sentiment score ζ_d of a document d can thus range from -1.5 to 1.5 and can subsequently be used to determine the associated classification of intended sentiment c_d.

2.3.2 Sentiment Mappings

In today's Web, reviewers can often quantify their overall verdict by assigning stars (typically with a maximum of five) to their reviews. As the use of such star ratings has become a wide-spread phenomenon across domains, languages, and cultures, we assume that consensus exists with respect to the meaning of each star rating class, thus rendering a five-star rating scale a universal classification method for intended sentiment. Star ratings capture the extent to which an author intends to convey positive sentiment and are defined on an ordinal scale, such that, e.g., a four-star text is considered to be more positive than a three-star text.

When modeling the relation between language-specific word-based sentiment scores ζ and universal star ratings c, we initially assume higher language-specific sentiment scores to be associated with higher star ratings, as people intending to convey rather positive sentiment would intuitively use rather positive words. As such, texts belonging to, e.g., the four-star class should have higher word-based sentiment scores than three-star texts. We thus map language-specific word-based sentiment scores to universal star ratings by means of a monotonically increasing step function.

We can thus construct language-specific *sentiment mappings* $M : \zeta_d \to c_d$, which translate the language-specific word-based sentiment score ζ_d of a document d into a universal star rating c_d. Each mapping covers five star segments, i.e., sets of sentiment-carrying texts that have the same number of stars assigned to them. These five segments are separated by four boundaries, the position of which is based on the sentiment scores of the texts in each segment.

An intuitive sentiment mapping is depicted in Figure 2.1. One could expect one-star and five-star texts to represent the extreme negative and positive cases, respectively, i.e., those covered by respective sentiment scores below -1 and above 1. The three-star class would intuitively be centered around a sentiment score of 0, indicating neutral or mixed sentiment. The two-star and four-star classes should then cover the remaining ranges of negative and positive sentiment scores, representing rather negative and positive conveyed sentiment, respectively.

Many alternative mappings may exist for, e.g., different domains or languages. Mappings may be skewed towards positive or negative sentiment scores or sentiment boundaries may be unequally spread across the full range of considered sentiment scores. The challenge is to find an optimal set of boundaries for each considered domain or language. The goal of this optimization process is to minimize the total costs κ_b associated with a given set of boundaries b. We define these costs as the sum of the number of misclassifications $\epsilon_{c_i}(b)$ in each individual sentiment class $c_i \in \{1, 2, 3, 4, 5\}$, given the set of boundaries b, i.e.,

$$\kappa_b = \sum_{i=1}^{5} \epsilon_{c_i}(b) \,. \tag{2.2}$$

As such, the optimization yields a set of boundaries associated with the least possible number of misclassifications, subject to the constraint that the boundaries must be non-overlapping and ordered, while being larger than the sentiment score lower bound ζ_l (e.g., -1.5) and smaller than the sentiment score upper bound ζ_u (e.g., 1.5), i.e.,

$$\zeta_l < b_1 < b_2 < b_3 < b_4 < \zeta_u. \tag{2.3}$$

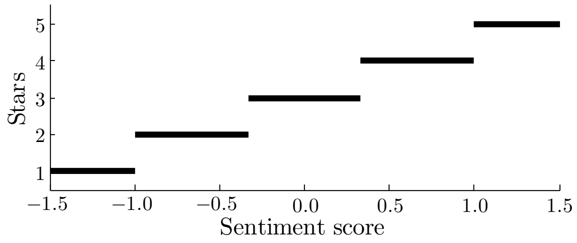


Figure 2.1: Intuitive mapping from sentiment scores to universal star ratings.

Finding an optimal set of sentiment boundaries is not trivial, as many combinations exist. Moreover, the sentiment boundaries are interdependent. Once an arbitrary boundary is set, it affects the possible locations of the other boundaries. Furthermore, classes may not be perfectly separable in the sole dimension of sentiment scores. Many algorithms can be used in order to cope with such issues. One could consider using a greedy algorithm when constructing a set of boundaries. Alternatively, heuristic or randomized optimization techniques like genetic algorithms may be applied in order to explore the solution space. Last, if the size of the data set allows, a brute force approach can be applied in order to assess all possible boundary sets at a certain level of granularity.

By using our proposed method, the sentiment conveyed by people's utterances of opinions in natural language can first be accurately analyzed by means of sentiment analysis tools tailored to the language used in these texts. The sentiment scores thus obtained can subsequently be transformed into star ratings by means of language-specific sentiment mappings, such that information monitoring systems can base their analyses on these universal classifications of intended sentiment rather than on less meaningful language-specific sentiment scores.

2.3.3 Star Ratings as a Non-Monotonic Function of Sentiment

The sentiment mapping method as proposed in Section 2.3.2 assumes that higher language-specific sentiment scores are associated with higher star ratings and that the sentiment classes associated with these star ratings are perfectly separable by non-overlapping boundaries in the dimension of language-specific sentiment scores. However, sentiment scores and star ratings may not be perfectly positively correlated, as, e.g., people tend to use rather positive words in negative reviews (Taboada et al., 2008). Moreover, sentiment classes may not be perfectly linearly separable in the dimension of language-specific sentiment scores. These concerns are not accounted for when modeling the relation between conveyed sentiment and intended sentiment as a monotonically increasing step function.

In this light, we propose to relax some constraints imposed on the model by our assumptions, such that the mapping $M : \zeta_d \rightarrow c_d$ between the sentiment scores ζ_d and star ratings c_d of a document d can possibly be more accurate. We propose to drop the monotonicity constraint and to allow for a non-linear relation between sentiment scores and star ratings, while not enforcing all star ratings to be represented in the mapping. To this end, we consider modeling our sentiment mappings by means of two machine learning approaches that are commonly used in state-of-the-art sentiment analysis methods, i.e., a naive Bayes model and a support vector machine (Pang and Lee, 2008).

2.3.4 Incorporating a Bag-of-Sentiwords Representation

The machine learning methods for modeling sentiment mappings $M : \zeta_d \rightarrow c_d$ proposed in Section 2.3.3 essentially represent sentiment-carrying text of a document d by means of a vector consisting of only one feature, i.e., the overall sentiment ζ_d conveyed by the natural language text as a whole. As our considered sentiment classes c_d may not be perfectly separable in the sole dimension of language-specific sentiment scores ζ_d, additional features may help improve the performance of the naive Bayes and support vector machine methods proposed in Section 2.3.3. The purpose of such additional features is to capture distinguishing characteristics of natural language content, such that the associated sentiment classes can be separated more accurately.

Sentiment-carrying words are considered to play a major role in conveying the overall sentiment of a text (Hogenboom et al., 2012b), as opinionated texts have been shown to significantly differ from non-opinionated texts in terms of occurrences of sentiment-carrying words (van der Meer et al., 2011). As such, sentiment-carrying words are attractive features to be included in vector representations of sentiment-carrying text, along with the overall sentiment conveyed by the text as a whole.

To this end, we propose to incorporate the *bag-of-sentiwords* representation proposed by Hogenboom et al. (2012b), thus introducing the occurrence of lexical representations of sentiment-carrying words retrieved from a sentiment lexicon as features in our vector representation of text. As such, we propose a sentiment mapping $M : (\zeta_d, \Xi) \rightarrow c_d$ of ζ_d to c_d, dependent on a vector of *bag-of-sentiwords* features Ξ. Following Hogenboom et al. (2012b), we opt for a binary representation of our additional features Ξ. In addition to incorporating *bag-of-sentiwords* features in our vectors, we propose to account for negation by differentiating between sentiment-carrying words and their negated counterparts, as accounting for negation has been shown to improve the performance of sentiment analysis approaches (Heerschop et al., 2011c; Hogenboom et al., 2011a).

In order to illustrate our vector representation of natural language text, let us consider the very negative sentence "*I would not recommend seeing that awful movie; it's just awful!*", which could be assigned a sentiment score of -1.5 and contains the negated, positive word "*recommend*" and two occurrences of the negative word "*awful*". This sentence could be represented as a vector $(-1.5, 0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0)$, with the first feature representing the sentiment score, the ones representing the occurrence of the negation of "*recommend*" and the occurrence of *awful*, and all zeros representing the occurrence of all other considered (possibly negated) sentiment-carrying words. This vector representation can be used for classifying the star rating of the associated text, while accounting for both its conveyed sentiment and the specific (negated and non-negated) words that convey this sentiment.

2.4 Evaluation

The methods proposed in Section 2.3 can be used to explore how language-specific sentiment scores can be converted into universal star ratings and how such mappings differ across collections of documents in different languages. In this section, we present our experimental set-up and discuss our experimental results.

2.4.1 Experimental Setup

In our analysis, we consider two distinct sets of similar documents. Our first data set consists of 1,759 short movie reviews in Dutch, crawled from various Web sites (Korte Reviews, 2011; Lemaire, 2011). The second data set considered in our current work consists of 46,315 short movie reviews in English (Metacritic, 2011; Short Reviews, 2011). These data sets essentially represent two distinct scenarios in which we assess our methods for mapping sentiment scores to star ratings. As we mainly assess these two scenarios in isolation, our analysis is not much affected by the difference in sample size of the two considered data sets.

Each review in our collections has been rated by its respective author on a scale of one to five or, for some Web sites, ten stars, with more stars implying a more positive verdict. We have constructed a ground truth on intended sentiment of our documents based on the ratings as given by their respective authors, where we have converted all scores on a ten-star scale to a five-star scale by dividing these scores by two and rounding the resulting scores to the nearest integer. For both considered languages, this process has yielded a data set in which the documents are distributed as shown in Figure 2.2.

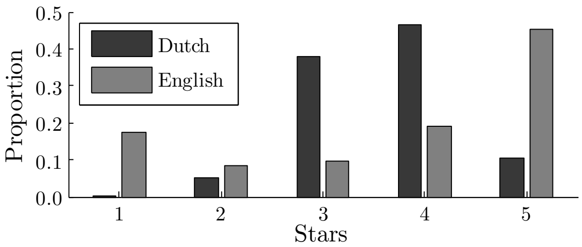


Figure 2.2: Distribution of documents over universal star rating classes, for our collections of Dutch and English movie reviews.

The documents in both data sets are first analyzed for the sentiment conveyed by the words used in their natural language content by means of an existing framework for lexicon-based sentiment analysis in multiple languages, i.e., Dutch and English (Bal et al., 2011). This framework is essentially a pipeline in which each component fulfils a specific task in analyzing the sentiment of a document. For each supported language, the sentiment analysis framework performs the text cleaning, word tokenization, POS tagging, word type classification, and sentiment scoring tasks as described in Section 2.3.1 by means of proprietary components in a C# implementation of the framework. The components for Dutch and English sentiment analysis are similar and use proprietary sentiment lexicons, which have been manually created and maintained.

When we use this existing lexicon-based sentiment analysis framework in order to score each document in our considered Dutch and English data sets for the sentiment conveyed by its natural language content, we obtain a set of 1,759 two-dimensional data points for Dutch and 46,315 similar two-dimensional data points for English. Each of these data points represents a paired observation of a language-specific sentiment score and the associated universal star rating of intended sentiment. For both considered languages, the data points thus obtained can be used to construct mappings between sentiment scores and star ratings for each considered language by means of the methods described in Section 2.3.

First, we consider modeling the relation between language-specific word-based sentiment scores and universal star ratings by means of a monotonically increasing step function (MIS). The goal here is to create a sentiment map similar to the intuitive one depicted in Figure 2.1. In order to optimize the location of the sentiment boundaries in these mappings, we use a brute force approach, where we optimize the performance of the resulting mapping in terms of number of misclassifications for all possible combinations of boundaries. We utilize a step size of 0.1, as this granularity renders a exploration of the full solution space feasible.

Second, we consider modeling our sentiment mappings by means of two machine learning (ML) methods that are commonly used in state-of-the-art sentiment analysis approaches (Hogenboom et al., 2012b), i.e., a naive Bayes (NB) model and a support vector machine (SVM). As such, we drop the monotonicity constraint of our first sentiment mapping method and allow for a non-linear relation between language-specific sentiment scores (SS) and (some) star ratings. In order to do so, we essentially represent each document by means of a vector consisting of only one feature, i.e., the overall sentiment conveyed by the natural language content of the document as a whole. We use existing WEKA (Hall et al., 2009) implementations of NB and SVM models, i.e., *NaiveBayes* and *SMO*, respectively, with their default settings.

Last, we expand the vector representations used by our NB and SVM models by incorporating *bag-of-sentiwords* (BoS) features in these ML methods. To this end, we introduce binary features into our vector representations of documents, signalling the presence of lexical representations of sentiment-carrying words retrieved from the proprietary general purpose sentiment lexicon used by our employed sentiment analysis framework. We only represent those lexical representations of sentiment-carrying words that occur in at least one of our documents. The presence of the negated counterparts of these sentiment-carrying words (if any) is signalled by separate features in order to account for negation. Negation is detected by our employed sentiment analysis framework, which considers a sentiment-carrying word to be negated if it is preceded by a negating modifier. We thus obtain 509 features representing the occurrence of (negated and non-negated) sentiment-carrying words in our Dutch documents and 884 features representing the occurrence of (negated and non-negated) sentiment-carrying words in our English documents.

Our models' quality is assessed by means of the 10-fold cross-validated overall accuracy and the macro-level F_1-measure. Accuracy is the overall proportion of correctly classified documents. The macro-level F_1-measure is the average F_1-measure of the individual star rating classes, weighted for their respective relative frequencies. An F_1-measure of a class is the harmonic mean of the precision and recall of that class. Precision quantifies the number of documents assigned to a class, relative to the number of documents that should have been assigned to that class, whereas recall quantifies the number of documents correctly assigned to that class, relative to the number of documents in that class. In our performance evaluation, we consider an absolute baseline of random classification. Because our methods may be biased towards overrepresented classes, the random baseline uses probability distributions that are equal to the class distributions as depicted in Figure 2.2. We assess the statistical significance of performance differences by means of a paired two-sample two-tailed t-test.

2.4.2 Experimental Results on Dutch Documents

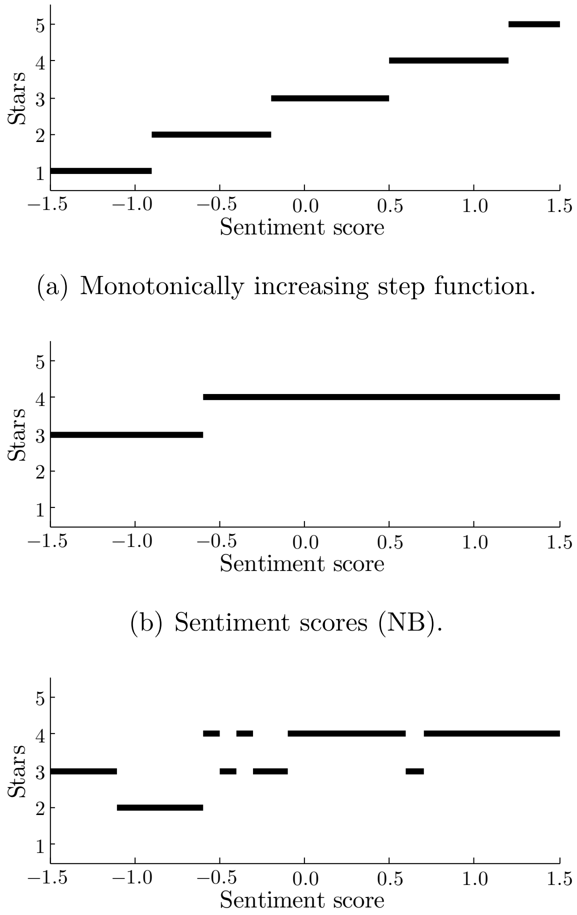
Our considered methods for mapping scores for language-specific sentiment conveyed by natural language text to intended sentiment, captured by universal star ratings, result in clearly distinct sentiment mappings. Examples of our constructed sentiment mappings for Dutch documents are visualized in Figure 2.3. Several observations can be made in these visualizations.

First, Figure 2.3(a) shows that the MIS model for our considered movie reviews in Dutch is more or less consistent with the intuitive sentiment mapping depicted in Figure 2.1. The classes of intended sentiment are approximately equally spread across the dimension of language-specific sentiment scores. Moreover, extreme sentiment scores are associated with extreme star ratings, whereas the three-star class, representing neutral or mixed sentiment, is associated with rather moderate sentiment scores.

Additionally, Figure 2.3(b) demonstrates that the SS models produce monotonically increasing step functions, even though this is not enforced. The sentiment mappings differ from the MIS mapping in that the SS models tend to focus on the distinction between three and four stars. This suggests that in our Dutch corpus, the overall sentiment conveyed by the words of a text as a whole can best be used for making a rough distinction between the two most frequent classes of intended sentiment, i.e., neutral or mixed sentiment and positive sentiment. A more fine-grained distinction between sentiment classes appears to be difficult when only considering the sentiment conveyed by the words in a text.

When taking into account the occurrence of specific sentiment-carrying words, such a more fine-grained distinction can be made, as the BoS models for Dutch documents typically show a more scattered, non-monotonic mapping from language-specific sentiment scores to star ratings, as visualized in Figure 2.3(c). This suggests that the relation between a text's conveyed and intended sentiment depends on the specific sentiment-carrying words used in the text.

Our distinct models perform significantly different from one another when using the constructed sentiment mappings for classifying text into one out of five star categories, as demonstrated by Tables 2.1, 2.2, and 2.3. The initial MIS sentiment mapping, albeit the most intuitive and interpretable model, is the worst performing sentiment mapping, being outperformed even by random classification. For our collection of Dutch movie reviews, the constraints of monotonicity, linearity, and representation of all star rating classes clearly thwart the predictive power that language-specific sentiment scores conveyed by words have with respect to the intended sentiment. The less constrained SS and BoS models significantly outperform the MIS model in terms of both macro-level F_1-score and overall accuracy, with performance improvements of more than 100%.



(a) Monotonically increasing step function.

(b) Sentiment scores (NB).

(c) Enriched with *bag-of-sentiwords* (SVM).

Figure 2.3: Typical sentiment mappings constructed on the training set of one specific fold of our evaluation on our Dutch corpus. These sentiment mappings depict how the (majority of) our considered opinionated Dutch documents are classified by means of our initial monotonically increasing step function (a), the best performing machine learning model using only the sentiment score as feature (b), and the best performing machine learning model additionally incorporating *bag-of-sentiwords* features (c).

Overall, on our Dutch corpus, the best performing models are the BoS models, which do not significantly differ from one another in terms of performance. Besides significantly outperforming the initial MIS model, the BoS models perform significantly better than the SS models and moreover exhibit a more consistent performance across sentiment classes, as signalled by their relatively low standard deviations. This indicates that for our Dutch corpus, sentiment conveyed by a text can be better mapped to universal star ratings of intended sentiment when accounting for the specific sentiment-carrying words used. Nevertheless, even our best models can map conveyed sentiment to intended sentiment only to a limited extent, as our highest accuracy and macro-level F_1-scores are about 50%.

Mapping	F_1 (μ)	F_1 (σ)	Accuracy
Random	0.350	0.123	0.349
MIS	0.210	**0.094**	0.193
SS (NB)	0.421	0.218	0.488
SS (SVM)	0.396	0.206	0.456
BoS (NB)	0.479	0.198	**0.524**
BoS (SVM)	**0.490**	0.153	**0.524**

Table 2.1: The weighted mean (μ) and standard deviation (σ) of the macro-level F_1-scores and accuracy over all classes, as computed for our considered methods on our Dutch movie review corpus, based on 10-fold cross-validation. The best performance is printed in bold for each performance measure.

Benchmark	Random	MIS	SS (NB)	SS (SVM)	BoS (NB)	BoS (SVM)
Random	0.000	-0.400***	0.204**	0.132	0.370***	0.401***
MIS	0.667***	0.000	1.006***	0.887***	1.283***	1.335***
SS (NB)	-0.169**	-0.502***	0.000	-0.060	0.138**	0.164**
SS (SVM)	-0.116	-0.470***	0.063	0.000	0.210**	0.238***
BoS (NB)	-0.270***	-0.562***	-0.121**	-0.174**	0.000	0.023
BoS (SVM)	-0.286***	-0.572***	-0.141**	-0.192***	-0.022	0.000

Table 2.2: Relative differences of 10-fold cross-validated macro-level F_1-scores of our considered approaches, benchmarked against one another on our collection of Dutch movie reviews. Performance differences marked with * are statistically significant at $p < 0.01$, those marked with ** are significant at $p < 0.001$, and those marked with *** are significant at $p < 0.0001$.

Benchmark	Random	MIS	SS (NB)	SS (SVM)	BoS (NB)	BoS (SVM)
Random	0.000	-0.446***	0.399***	0.306***	0.502***	0.502***
MIS	0.805***	0.000	1.526***	1.357***	1.712***	1.712***
SS (NB)	-0.285***	-0.604***	0.000	-0.067*	0.074*	0.074
SS (SVM)	-0.234***	-0.576***	0.071*	0.000	0.150***	0.150***
BoS (NB)	-0.334***	-0.631***	-0.069*	-0.131***	0.000	0.000
BoS (SVM)	-0.334***	-0.631***	-0.069	-0.131***	0.000	0.000

Table 2.3: Relative differences of the 10-fold cross-validated overall accuracy of our considered approaches, benchmarked against one another on our collection of Dutch movie reviews. Performance differences marked with * are statistically significant at $p < 0.01$, those marked with ** are significant at $p < 0.001$, and those marked with *** are significant at $p < 0.0001$.

2.4.3 Experimental Results on English Documents

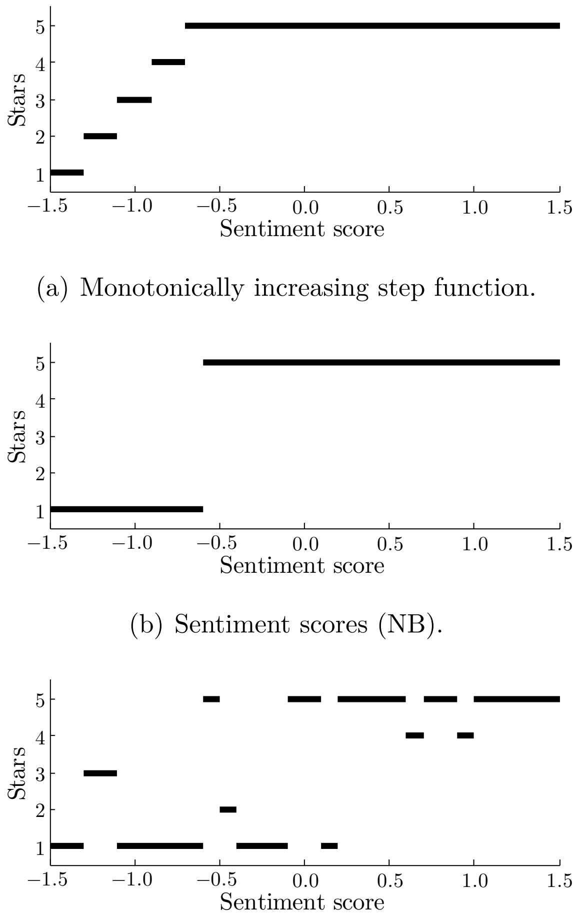
The sentiment mappings created using our methods on our collection of English movie reviews look different from the Dutch sentiment mappings presented in Section 2.4.2. Figure 2.4 exhibits some of the distinctive features of our constructed sentiment mappings for our data set of English documents.

First, the initial MIS model and, to a lesser extent, even the more sophisticated SS and BoS models show a bias towards negative sentiment scores. In general, in our corpus of English movie reviews, moderately positive and even moderately negative sentiment scores of sentiment conveyed by natural language content are typically already associated with the highest, i.e., most positive star ratings of intended sentiment by all of our models.

Additionally, similarly to our findings for our Dutch corpus, Figure 2.4(b) demonstrates that the SS models produce monotonically increasing step functions for our English corpus, even though this is not enforced. As is the case in our Dutch corpus, the English sentiment mappings produced by our SS models differ from the MIS sentiment mapping in that the SS models tend to focus on the distinction between only two star rating classes which are relatively frequent in the corpus. Our English SS models tend to focus on the distinction between clearly positive and clearly negative documents, i.e., those associated with five stars and one star, respectively. In our corpus, the overall sentiment conveyed by the words of an English text as a whole can apparently best be used for making a rough distinction between two classes of intended sentiment, i.e., positive and negative sentiment.

A more fine-grained distinction between classes of intended sentiment is possible when not only considering the sentiment conveyed by the words in a text, but by additionally accounting for the specific words conveying this sentiment as well. The BoS models for our collection of English movie reviews typically show a rather scattered, non-monotonic mapping from conveyed sentiment to intended sentiment. The bias towards negative sentiment scores is less apparent in the BoS models than it is in the other models and, overall, the relation between conveyed sentiment and intended sentiment appears to be more complex. For instance, some documents with overall positive scores are classified as having the most negative rating when accounting for the distinct sentiment-carrying words used. Similarly, some documents with overall negative scores are classified as having the most positive star rating by our BoS-based sentiment mappings. As such, the specific sentiment-carrying words used in English text appear to play an important role in the relation between conveyed and intended sentiment.

Tables 2.4, 2.5, and 2.6 show that the MIS and SS (SVM) models are the worst performing models on our English corpus, hardly outperforming random classification of intended sentiment. The SS (NB) approach yields somewhat better sentiment mappings.



(a) Monotonically increasing step function.

(b) Sentiment scores (NB).

(c) Enriched with *bag-of-sentiwords* (SVM).

Figure 2.4: Typical sentiment mappings constructed on the training set of one specific fold of our evaluation on our English corpus. These sentiment mappings depict how the (majority of) our considered opinionated English documents are classified by means of our initial monotonically increasing step function (a), the best performing machine learning model using only the sentiment score as feature (b), and the best performing machine learning model additionally incorporating *bag-of-sentiwords* features (c).

Yet, it is only when the occurrence of sentiment-carrying words is taken into account that performance clearly improves with over 50% in terms of macro-level F_1-scores, and over 10% in terms of overall accuracy, compared to the MIS model. Compared to our findings on our Dutch corpus, these improvements are relatively small, yet significant. This is caused by our English MIS model being of a comparatively good quality in terms of performance, as opposed to our Dutch MIS model. Our best English sentiment mappings exhibit an overall accuracy of over 50%, yet the best macro-level F_1-score is approximately 45%. As such, for our English corpus, the scores of conveyed sentiment can be mapped to star ratings of intended sentiment only to a limited extent.

Mapping	F_1 (μ)	F_1 (σ)	Accuracy
Random	0.305	0.292	0.449
MIS	0.296	0.303	0.457
SS (NB)	0.357	0.293	0.480
SS (SVM)	0.284	0.311	0.454
BoS (NB)	**0.437**	**0.241**	0.505
BoS (SVM)	0.425	0.257	**0.511**

Table 2.4: The weighted mean (μ) and standard deviation (σ) of the macro-level F_1-scores and accuracy over all classes, as computed for our considered approaches on our English movie reviews, based on 10-fold cross-validation. The best performance is printed in bold for each performance measure.

Benchmark	Random	MIS	SS (NB)	SS (SVM)	BoS (NB)	BoS (SVM)
Random	0.000	-0.031***	0.169***	-0.071***	0.429***	0.391***
MIS	0.032***	0.000	0.207***	-0.042***	0.475***	0.435***
SS (NB)	-0.145***	-0.171***	0.000	-0.206***	0.222***	0.189***
SS (SVM)	0.076***	0.043***	0.259***	0.000	0.538***	0.497***
BoS (NB)	-0.300***	-0.322***	-0.182***	-0.350***	0.000	-0.027*
BoS (SVM)	-0.281***	-0.303***	-0.159***	-0.332***	0.027*	0.000

Table 2.5: Relative differences of 10-fold cross-validated macro-level F_1-scores of our considered approaches, benchmarked against one another on our collection of English movie reviews. Performance differences marked with * are statistically significant at $p < 0.01$, those marked with ** are significant at $p < 0.001$, and those marked with *** are significant at $p < 0.0001$.

Benchmark	Random	MIS	SS (NB)	SS (SVM)	BoS (NB)	BoS (SVM)
Random	0.000	0.018***	0.070***	0.012***	0.125***	0.139***
MIS	-0.018***	0.000	0.051***	-0.006***	0.105***	0.119***
SS (NB)	-0.065***	-0.048***	0.000	-0.054***	0.052***	0.065***
SS (SVM)	-0.012***	0.006***	0.057***	0.000	0.112***	0.125***
BoS (NB)	-0.111***	-0.095***	-0.049***	-0.100***	0.000	0.012
BoS (SVM)	-0.122***	-0.106***	-0.061***	-0.111***	-0.012	0.000

Table 2.6: Relative differences of the 10-fold cross-validated overall accuracy of our considered approaches, benchmarked against one another on our collection of English movie reviews. Performance differences marked with * are statistically significant at $p < 0.01$, those marked with ** are significant at $p < 0.001$, and those marked with *** are significant at $p < 0.0001$.

2.4.4 Overall Experimental Results

The results presented in Sections 2.4.2 and 2.4.3 demonstrate that sentiment mappings may have different characteristics for distinct collections of documents – in our case documents written in different languages. The constructed sentiment mappings for our considered data sets do however exhibit some similar patterns. First, in both data sets, allowing for a non-linear, non-monotonic relation between conveyed sentiment and intended sentiment – possibly not covering all star rating classes – yields sentiment mappings that tend to make only crude distinctions between sentiment classes, e.g., positive and negative. Second, the relation between conveyed and intended sentiment only partly depends on the specific sentiment-carrying words used in a document's natural language content. Involving the occurrence of such words in the analysis of a document and its associated intended sentiment enables a significantly better distinction between the five considered classes of intended sentiment, as compared to not accounting for such features. In this light, methods for classifying intended sentiment would benefit from a type of analysis in which the specifics of natural language content are taken into account.

Our best performing sentiment mappings for both Dutch and English documents tend to be non-monotonic – occasionally, more positive sentiment scores are associated with a more negative intended sentiment, and vice versa. As our considered sentiment scores are mainly constituted by sentiment-carrying words and their modifiers, and largely ignore semantic and structural aspects of content, our results suggest that intended sentiment may not necessarily be conveyed by the sentiment-carrying words and their modifiers per se, but rather by the way in which these words are used. An additional explanation for the observed non-monotonicity of the constructed sentiment mappings lies in our mappings only covering two dimensions of the sentiment analysis problem, i.e., the dimension of language-specific sentiment scores and the dimension of intended sentiment. Other factors such as (the semantics of) the sentiment-carrying words used, rhetorical aspects, or even cultural aspects may be affecting the relation between sentiment scores and intended sentiment, thus yielding non-monotonic mappings in the two considered dimensions. As such factors are not explicitly accounted for in our mappings, we can only successfully map sentiment conveyed by natural language text to intended sentiment to a limited extent.

An error analysis of our results on both collections of Dutch and English texts reveals that many classification errors are indeed caused by our sentiment analysis methods accounting for *what* is said rather than for *how* sentiment-carrying words are used. For instance, people tend to discuss different aspects of a movie and possibly even of other movies before arriving at their conclusions. The sentiment conveyed by the conclusions in such movie reviews appears to be a better proxy for the overall intended sentiment.

Another compelling illustrative example from our data is a review in which the author heavily cites and criticizes negative reviews and, by doing so, in fact conveys a positive opinion while almost exclusively using negative sentiment-carrying words. Even our best methods classify the intended sentiment of this text as very negative, i.e., one star, whereas it should have been assigned five stars. In this light, accounting for semantics or the (rhetorical) role of sentiment-carrying words in a text may help improve our mappings of conveyed sentiment to intended sentiment.

2.5 Conclusions

In this chapter, we have proposed to use language-specific word-based sentiment scores in order to classify natural language text into universal star ratings that capture people's intended sentiment. We envisage such mappings to be useful in analytical tools for quantifying people's true sentiment, independent of domain, language, or culture. The results of our experiments with respect to modeling the relation between conveyed and intended sentiment for Dutch and English corpora suggest that the way natural language reveals people's intended sentiment may differ across collections of documents. Additionally, the relation between conveyed and intended sentiment of documents in both considered data sets only partly depends on the sentiment scores of the words in a document. When accounting for the occurrence of specific sentiment-carrying words used in a document's natural language content, our results show that language-specific sentiment scores can separate universal classes of intended sentiment to a better, but still limited extent.

The findings presented in this chapter indicate that, in practice, language-specific word-based sentiment scores form a good starting point for capturing people's truly intended sentiment, especially when combined with information on which specific sentiment-carrying words constitute these scores. However, in order to be able to more accurately capture people's intended sentiment by means of analyzing the natural language used by people to convey their sentiment, more aspects of content, other than just the sentiment-carrying words used, may need to be taken into account.

Therefore, the following chapters explore the viability of exploiting other aspects of text when analyzing people's intended sentiment. The key may not be so much in *what* people say, but rather in *how* they use sentiment-carrying words in their motivation for their opinion. In this light, the next chapters deal with the use of additional explicit and latent cues in order to better understand the sentiment that people intend to convey. Emoticons could be useful explicit proxies for an author's intended sentiment, whereas semantics or the roles that words play in a text may be useful latent cues.

Chapter 3

Emoticon-Guided Polarity Classification of Text*

SINCE *people increasingly use emoticons in written text on the Web in order to express, emphasize, or disambiguate their sentiment, it is crucial for automated sentiment analysis tools to correctly account for such graphical cues for sentiment. We analyze how emoticons typically convey sentiment by interacting with the (sentiment conveyed by) words in a text on a semantic level. Additionally, we propose and evaluate a novel method for exploiting this mechanism with a manually created emoticon sentiment lexicon in a lexicon-based polarity classification method. We evaluate our approach on 2,080 Dutch tweets and forum messages, which all contain emoticons. We validate our findings on 10,069 English reviews of apps, some of which contain emoticons. We find that accounting for the sentiment conveyed by emoticons on a paragraph level – and, to a lesser extent, on a sentence level – significantly improves polarity classification performance. Whenever emoticons are used, their associated sentiment tends to dominate the sentiment conveyed by textual cues and thus forms a good proxy for the polarity of text.*

*This chapter is based on the article "A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak. Exploiting Emoticons in Polarity Classification of Text. *Journal of Web Engineering*, 14(1):22–40, 2015."

3.1 Introduction

Nowadays, popular Web sites like Twitter, Blogger, and Epinions allow their users to vent opinions on just about anything through an ever-increasing amount of short messages, blog posts, or reviews. This social interaction through the Web, i.e., the Social Web, yields a continuous flow of an overwhelming amount of data, containing traces of valuable information – people's sentiment with respect to products, brands, etcetera. The abundance of user-generated content published through the Social Web renders automated information monitoring tools crucial for today's businesses (Ojokoh and Kayode, 2012).

Automated sentiment analysis techniques come to answer this need (Cambria et al., 2013; Feldman, 2013; Liu, 2012; Pang and Lee, 2008). Sentiment analysis refers to a broad area of natural language processing, computational linguistics, and text mining. Typically, the goal is to determine the polarity of natural language texts. An intuitive approach would involve scanning a text for cues signaling its polarity. Typical cues in written text could be sentiment-carrying words.

In face-to-face communication, sentiment can often be deduced from visual cues like frowning or smiling. However, in plain-text computer-mediated communication, such visual cues are lost. Over the years, people have embraced the usage of so-called emoticons as an alternative to face-to-face visual cues in (on-line) computer-mediated communication like virtual utterances of opinions in the Social Web. In this light, we define emoticons as visual cues used in texts to replace normal visual cues like frowning or smiling in order to express, emphasize, or disambiguate one's sentiment. Emoticons are typically sequences of typographical symbols such as “*:*”, “*=*”, “*-*”, “*)*”, or “*(*” and commonly represent facial expressions. Emoticons can be read either sideways, like “*:-(*” (a sad face), or normally, like “*(^_^)*” (a happy face).

Several types of automated polarity classification methods have been proposed in recent years (Cambria et al., 2013; Feldman, 2013; Liu, 2012; Pang and Lee, 2008). Many state-of-the-art methods represent natural language text as an unordered collection of the words occurring in the text. This allows for vector representations of text, enabling the use of machine learning techniques for classifying its polarity. Features in such representations may be, e.g., words or parts of words. However, machine learning polarity classifiers typically require a lot of training data in order to function properly. Moreover, even though machine learning classifiers may perform very well in the domain that they have been trained on, their performance drops significantly when they are used in another domain (Taboada et al., 2008). Additionally, machine learning polarity classifiers typically give little insight into why a text is assigned a specific polarity classification.

In this light, alternative lexicon-based methods have recently gained (renewed) attention (Cesarano et al., 2006; Devitt and Ahmad, 2007; Ding et al., 2008; Heerschop et al., 2011a,c; Taboada et al., 2011). These essentially rule-based methods use sentiment lexicons for retrieving the polarity of individual words and subsequently aggregate these scores – e.g., by computing a (weighted) sum or average of the individual word scores – in order to determine a text's overall polarity. A sentiment lexicon typically contains words and their associated sentiment, possibly differentiated by Part-of-Speech (POS) and/or meaning (Baccianella et al., 2010).

Recent findings suggest that people's intended sentiment is not so much conveyed by the words in a text per se (see also Chapter 2), but rather by how these words are used (Bal et al., 2011; Chenlo et al., 2013; Heerschop et al., 2011a). Emoticons could be helpful here, as an emoticon may for instance signal the intended sentiment of an otherwise objective statement, e.g., "*This product does not work :-(*". However, today's lexicon-based polarity classification approaches do not typically consider the semantics of emoticons. Conversely, one of the first steps in most existing work is to remove many of the typographical symbols typically constituting emoticons, thus preventing emoticons from being detected at all. We aim to investigate how the sentiment conveyed by text depends on emoticons, and how we can exploit this in order to improve lexicon-based polarity classification.

The remainder of this chapter is structured as follows. First, Section 3.2 elaborates on how emoticons are used in computer-mediated communication, as well as on how emoticons have already been exploited in existing sentiment analysis approaches. Then, in Section 3.3, we analyze how emoticons are typically related to the polarity of the text they occur in and we additionally propose a method for harvesting information from emoticons when analyzing the polarity of text. The performance of our novel approach is assessed in Section 3.4. Last, we conclude in Section 3.5.

3.2 Related Work

Research has demonstrated that humans are influenced by the use of nonverbal cues in face-to-face communication (Childers and Houston, 1984; Shepard, 1967). Nonverbal cues have even been shown to dominate verbal cues in face-to-face communication in case verbal and nonverbal cues are equally strong (Burgoon et al., 1996). Nonverbal cues are clearly important for people's understanding of the intentions and emotions of whomever they communicate with. Hence, translating these findings to computer-mediated communication does not appear to be too far-fetched, if it were not for the fact that plain-text computer-mediated communication does not leave much room for nonverbal cues.

However, users of computer-mediated communication have found their ways of overcoming the lack of personal contact by using emoticons. The first emoticon was used on September 19, 1982 by professor Scott Fahlman in a message on the computer science bulletin board of Carnegie Mellon University. In his message, Fahlman proposed to use the character sequences "*:-)*" and "*:-(*" in order to clearly distinguish jokes from more serious matters, respectively. It did not take long before the phenomenon of emoticons had spread to a much larger community. People started sending yells, hugs, and kisses by using graphical symbols formed by characters found on a typical keyboard. A decade later, emoticons had found their way into everyday computer-mediated communication and had become the paralanguage of the Web (Marvin, 1995). By then, 6% of the messages on electronic mailing lists (Rezabek and Cochenour, 1998) and 13% of UseNet newsgroup posts (Witmer and Katzman, 1997) were estimated to contain emoticons.

Thus, nonverbal cues have emerged in computer-mediated communication. It should however be noted that these nonverbal cues in computer-mediated communication are conceptually different from nonverbal cues in face-to-face communication. Real-life cues like laughing and weeping are often considered to be involuntary ways of expressing oneself in face-to-face communication, whereas the use of their respective equivalents "*:-)*" and "*:-(*" in computer-mediated communication is intentional (Kendon, 1987). As such, emoticons enable people to indicate subtle mood changes, to signal irony, sarcasm, and jokes, and to express, emphasize, or disambiguate their (intended) sentiment, perhaps even more than nonverbal cues in face-to-face communication can. Therefore, harvesting information from emoticons appears to be a viable strategy to improve the state-of-the-art of sentiment analysis. Yet, the question is not so much *whether*, but rather *how* we should account for emoticons when classifying the polarity of a document.

Even though recent lexicon-based polarity classification approaches explore promising new directions of incorporating structural and semantic aspects of content (Heerschop et al., 2011a), they typically fail to harvest information from potentially important cues for sentiment in today's user-generated content – emoticons. Nevertheless, emoticons have already been exploited to a limited extent, mainly for automated data annotation. For instance, a crude distinction between a handful of positive and negative emoticons has been used in order to automatically generate data sets with positive and negative samples of text (Read, 2005). The results of Read (2005) suggest that the sentiment conveyed by emoticons is topic- and domain-independent. These findings have been successfully applied in later work in order to automatically construct sets of positive and negative tweets (Davidov et al., 2010; Pak and Paroubek, 2010), or collections of tweets in alternative sentiment categories, such as angry and sad emotional states (Zhao et al., 2012).

Combining such automatically annotated training data with manually labeled training data has been shown to yield sentiment classifiers that outperform similar classifiers that have been trained on manually annotated training data only (Liu et al., 2012).

In other work, a few emoticons have been used as features for polarity classification, in addition to more common features such as sentiment-carrying words (Thelwall et al., 2010). However, the results of the latter work do not indicate that a significant polarity classification performance improvement over ignoring emoticons can be achieved when treating emoticons as if they are normal sentiment-carrying words that do not interact with other cues on a semantic level. Provided that emoticons are, nevertheless, important cues for sentiment in today's user-generated content, the key to harvesting information from emoticons lies in understanding how they relate to a text's overall polarity.

Yet, existing work does not focus on investigating how emoticons affect the polarity of natural language text, nor on exploring how this mechanism can be exploited in lexicon-based polarity classification. In the remainder of this chapter, we address this hiatus.

3.3 Emoticons and Polarity

In order to exploit emoticons in an automated polarity classification setting, we first need to analyze how emoticons are typically related to the polarity of the text they occur in. Insights into what parts of a text are affected by emoticons in which way are crucial enablers for successful polarity classifiers that harvest information from emoticons.

3.3.1 Emoticons as Cues for Polarity

We have performed a qualitative analysis of a collection of 2,080 Dutch tweets and forum messages, in order to assess the role of emoticons in conveying the sentiment of a text. This content has been randomly sampled from search results from Twitter and Google discussion groups, when querying for brands like Vodafone, KLM, and Kinect.

The first hypothesis that we have evaluated on our data is the hypothesis of emoticons having a rather local effect, i.e., emoticons affecting a paragraph or a sentence. Paragraphs typically address different points of view for a single topic or different topics, thus rendering the applicability of an emoticon in one paragraph to another paragraph rather unlikely. In our sample collection, upon inspection, emoticons generally have a paragraph-level effect for those paragraphs containing only one emoticon. In case a paragraph contains multiple emoticons, the analysis of our sample shows that an emoticon is generally more likely to affect the sentence in which it occurs.

Sentence	How	Sentiment
I love my work :-D	Intensification	Positive
The movie was bad :-D	Negation	Positive
:-D I got a promotion	Only sentiment	Positive
-_- *I love my work*	Negation	Negative
The movie was bad -_-	Intensification	Negative
I got a promotion -_-	Only sentiment	Negative

Table 3.1: Typical examples of how emoticons can be used to convey sentiment.

In our sample, 84% of all emoticons are placed at the end of a paragraph, 9% are positioned somewhere in the middle of a paragraph, and 7% are used at the beginning of a paragraph. This positioning of emoticons suggests that it is typically not a single word, but rather a text segment that is affected by an emoticon. Additionally, these results imply that in case an emoticon is used in the middle of a paragraph with multiple emoticons, the emoticon is statistically more likely to be associated with the preceding text segment.

In addition to assessing *what* is affected by emoticons, we have analyzed *how* emoticons affect text as well. Our sample shows that emoticons can generally be used in three ways. First, emoticons can be used to express sentiment when sentiment is not conveyed by any clearly positive or negative words in a text segment, thus rendering the emoticons to be carrying the only sentiment in such cases. Second, emoticons can emphasize sentiment by intensifying the sentiment already conveyed by sentiment-carrying words. Third, emoticons can be used to disambiguate sentiment, for instance in cases where the sentiment associated with sentiment-carrying words needs to be negated. Some illustrative examples can be found in Table 3.1.

Table 3.1 clearly demonstrates how the sentiment associated with a piece of text can differ when using different emoticons, i.e., the happy emoticon "*:-D*" and the "-_-" emoticon indicating extreme boredom or disagreement, irrespective of the emoticons' position in the text. The sentiment carried by an emoticon is independent from its embedding text, rendering word sense disambiguation techniques (Navigli, 2009) not useful for emoticons. Thus, the sentiment of emoticons appears to be dominating the sentiment carried by verbal cues in sentences, if any.

In some cases, this domination may be a crucial property of emoticons which can be exploited by automated sentiment analysis approaches. For instance, when an emoticon is the only sentiment-conveying cue in a sentence, we are typically dealing with a phenomenon that we refer to as factual sentiment. For example, the sentence "*I got a promotion*" does nothing more than objectively stating the fact that one was promoted.

Emoticon synset	Emoticons
Happiness	:-D, =D, xD, (^_^)
Sadness	:-(, =(
Crying	:'(, ='(, (;_;)
Boredom	-_-, -.-, (>_<)
Love	<3, (L)
Embarrassment	:-$, =$, >///<

Table 3.2: Typical examples of emoticon synsets.

However, getting a promotion is usually linked to a positive emotion like happiness or pride. Therefore, human interpreters could typically be inclined to acknowledge the implied sentiment and thus consider the factual statement to be a positive statement. This however requires an understanding of context and involves incorporating real-world knowledge into the process of sentiment analysis. For machines, this is a cumbersome task. In this light, emoticons can be valuable cues for deriving an author's intended sentiment.

3.3.2 Framework

We propose a novel framework for automated document-level polarity classification, which takes into account the semantics of emoticons. This framework detects emoticons, determines their sentiment, and assigns this sentiment to the text affected by the emoticons. The emoticon-based information thus obtained is then combined with the sentiment conveyed by verbal cues in the remaining unaffected text, in order to classify the polarity of a document as either positive or negative. Our framework, depicted in Figure 3.1, builds upon existing work (Bal et al., 2011) and is a pipeline in which each component fulfills a specific task in analyzing the sentiment of a document. Here, a document is a piece of text that can be as small as a one-line tweet or as big as a news article, review, blog, or forum message with multiple paragraphs, as long as it is one coherent piece of text.

First, we load a document to be analyzed for sentiment. Then, the document is split into text segments, which may be either paragraphs or sentences (step 1). Sentiment analysis is subsequently initially performed on segment level, after which the segment-level sentiment analysis results are combined.

Each text segment is subsequently checked for the presence of emoticons (step 2). To this end, we propose an emoticon sentiment lexicon, which we define as a list of character sequences – representing emoticons – and their associated sentiment scores. These emoticons may be organized into emoticon synsets, which we define as groups of emoticons denoting the same emotion. Table 3.2 shows examples of such emoticon synsets.

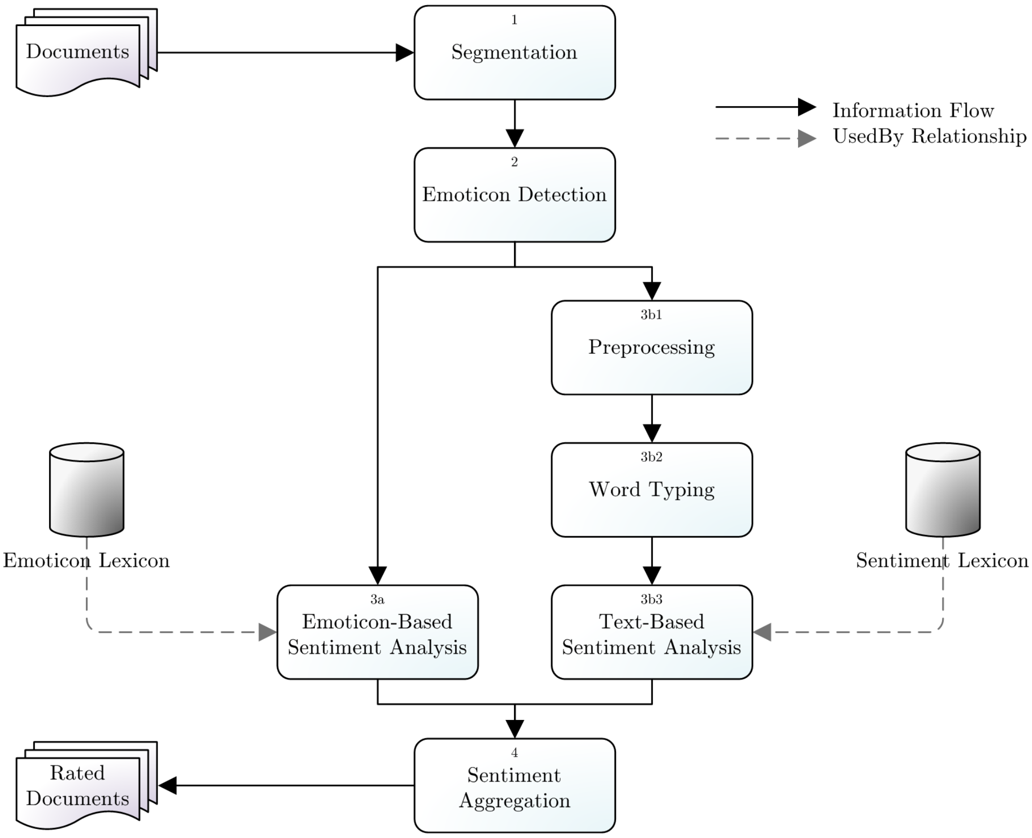


Figure 3.1: Overview of our emoticon-guided sentiment analysis framework.

When checking a text segment for the presence of emoticons, we compare each word in the segment with the emoticon sentiment lexicon. Here, we consider words to be character sequences, separated by whitespace characters. If a word in a text segment matches a character sequence in the emoticon sentiment lexicon, the segment is rated for sentiment based on the sentiment imposed onto the text by its emoticons (step 3a). Else, the segment is analyzed for the sentiment conveyed by its sentiment-carrying words (step 3b1–3).

In case a text segment is analyzed based on the emoticons it contains (step 3a), the segment is assigned a sentiment score equal to the total (summed) sentiment associated with all of its emoticons, as derived from the emoticon sentiment lexicon. Sentiment scores of sentiment-carrying words (if any) are ignored in this process, as the sentiment of emoticons tends to dominate the sentiment carried by verbal cues (see Section 3.3.1).

In order to analyze a text segment for the sentiment conveyed by its sentiment-carrying words (step 3b1–3), it is first preprocessed by removing diacritics and other special characters (step 3b1) and identifying each word's POS, lemma, and its purpose in the text, i.e., sentiment-carrying term, modifying term, or irrelevant term (step 3b2). We consider modifying terms to change the sentiment of corresponding sentiment-carrying word(s).

Negations are assumed to change the sentiment sign and amplifiers are assumed to multiply the sentiment of affected words with an appropriate factor. After determining the word types, the text segment is rated for its conveyed sentiment by means of a lexicon-based sentiment scoring method that computes the sentiment of the text segment as the total sentiment score of all (modified) sentiment-carrying words in the segment (step 3b3).

As such, the sentiment score ζ_{s_i} of the i-th segment s_i of document d can be computed as either a function of the sentiment scores ζ_{e_j} of each emoticon e_j in segment s_i, or as a function of the sentiment scores ζ_{t_j} of each sentiment-carrying word t_j and the weight w_{t_j} of its modifying term (if any, else, this weight defaults to 1), i.e.,

$$\zeta_{s_i} = \begin{cases} \sum_{j=1}^{V_i} \zeta_{e_j} & \text{if } V_i > 0, \\ \sum_{j=1}^{T_i} \left(\zeta_{t_j} \cdot w_{t_j}\right) & \text{otherwise,} \end{cases} \tag{3.1}$$

with V_i the number of visual cues for sentiment in segment s_i and T_i the number of sentiment-carrying textual cues (i.e., combinations of sentiment-carrying words and their modifiers) in the segment. In (3.1), ζ_{e_j} and ζ_{t_j} are real numbers ranging from -1 (negative) to 1 (positive).

After determining the sentiment conveyed by each individual text segment, all text segments are recombined into a single document. Note that a document can have both segments with and without emoticons. The sentiment score ζ_d of a document d is then calculated as the average over all segment-level sentiment scores, i.e.,

$$\zeta_d = \frac{\sum_{i=1}^{p} \zeta_{s_i}}{\sum_{i=1}^{p} \left(V_i + (a_i \cdot T_i)\right)}, \tag{3.2}$$

with p the number of segments of document d and a_i a Boolean variable indicating whether a full sentiment analysis needs to be performed on the textual cues of text segment s_i (1) or not (0), i.e.,

$$a_i = \begin{cases} 0 & \text{if } V_i > 0, \\ 1 & \text{otherwise.} \end{cases} \tag{3.3}$$

A negative document-level sentiment score thus computed typically indicates a negative polarity (-1), whereas other scores indicate a positive polarity (1). The classification c_d of document d is therefore defined as a function of its sentiment score ζ_d, i.e.,

$$c_d = \begin{cases} -1 & \text{if } \zeta_d < 0, \\ 1 & \text{otherwise.} \end{cases} \tag{3.4}$$

3.4 Polarity Classification by Exploiting Emoticons

By means of a set of experiments, we evaluate our novel method of polarity classification of natural language text by exploiting emoticons. The setup of our experiments is detailed in Section 3.4.1. We present our experimental results and our validation of these results in Sections 3.4.2 and 3.4.3, respectively. Last, we discuss some caveats with respect to our findings in Section 3.4.4.

3.4.1 Experimental Setup

For our current purpose, we evaluate the performance of an implementation of our method on a collection of Dutch tweets and forum posts. Additionally, we validate our findings on another data set that covers a different domain and language, i.e., a collection of English app reviews.

3.4.1.1 Data

Our collection of Dutch documents consists of 2,080 Dutch tweets and forum messages. We have randomly sampled these pieces of natural language text from search results from Twitter and Google discussion groups when querying for brands like Vodafone, KLM, Kinect, while making sure to select only those texts that contain emoticons. Three human annotators have manually annotated these documents for their associated polarity, i.e., positive or negative, until they reached agreement. The resulting data set consists of 1,067 positive documents and 1,013 negative documents. Emoticons occur in all documents in this data set.

The 10,069 English app reviews in our validation set have been crawled from Apple's App Store for the United Kingdom. We have collected reviews for various apps, ranging from, among others, the Dropbox, Gmail, and WhatsApp Messenger apps to the TomTom Europe, Bloomberg, and Pocket Whip apps. For each review, the associated sentiment had already been annotated by its author by means of a star rating, ranging from one star (very negative) to five stars (very positive). We have automatically converted these star ratings into binary polarity classifications, by assigning a negative classification to reviews with one, two, or three stars, and a positive classification to reviews with four or five stars. As a result, our collection consists of 7,017 positive and 3,052 negative English app reviews. By applying our emoticon detection method described in Section 3.3.2, we have automatically detected emoticons in a subset of 655 English app reviews, i.e., in 527 positive and 128 negative documents.

3.4.1.2 Implementation

In a C# implementation of our framework for polarity classification of Dutch documents, we look for empty lines or lines starting with an indentation in order to split a document into paragraphs. When splitting a document into sentences, we look for punctuation marks, such as “.”, “*!*”, and “*?*”, as well as for emoticons, as most emoticons are placed at the end of a text segment (see Section 3.3.1). We utilize a proprietary maximum-entropy based POS tagger for Dutch and a proprietary sentiment lexicon for Dutch words. This sentiment lexicon enables us to retrieve both the sentiment scores of sentiment-carrying words and the values for their associated modifiers, i.e., negators or amplifiers, if any.

In a similar, Java-based implementation of our framework for polarity classification of English documents, we detect paragraphs by making use of empty lines in a document, as these empty lines are used in our data to separate paragraphs from one another. Furthermore, we employ the Stanford Tokenizer (Manning et al., 2010) for identifying sentences and words in the identified paragraphs. For POS tagging and lemmatization of words, we use the OpenNLP (Baldridge and Morton, 2004) POS tagger and the Java WordNet Library (JWNL) API (Walenz and Didion, 2008), respectively. We use the identified lemma and POS of each word in order to retrieve its associated sentiment score from the SentiWordNet 3.0 (Baccianella et al., 2010) sentiment lexicon, which contains positivity, negativity, and objectivity scores for each entry. We use this information to compute sentiment scores for each word by subtracting the negativity score from the positivity score associated with its first (i.e., most common) sense, thus yielding a real number ranging from -1 (very negative) to 1 (very positive). Following recent findings (Hogenboom et al., 2011a), we account for negation by inverting the polarity of the two words following a negation keyword that is listed in a negation lexicon. Last, we account for amplification by means of an existing amplification lexicon, listing amplification keywords and their effect on the sentiment conveyed by the first succeeding word (Taboada et al., 2011).

3.4.1.3 Emoticon Sentiment Lexicon

One of the key elements in our novel emoticon-based polarity classification framework is the emoticon sentiment lexicon. Several lists of emoticons are readily available (ComputerUser, 2013; Gil, 2013; Marks, 2004; Marshall, 2003; Msgweb, 2006; Sharpened, 2013; Thelwall et al., 2011; Wikipedia, 2013). We propose to combine these eight existing lists into one large emoticon sentiment lexicon. In this process, we leave out duplicate entries. Additionally, we leave out character representations of body parts and representations of objects, as the latter two types of emoticons do not carry any sentiment.

This process yields a list of 477 emoticons representing facial expressions or body poses like thumbs up. Three human annotators have manually rated the emoticons in our lexicon for their associated sentiment, i.e., -1.0 (negative), 0.0 (neutral), or 1.0 (positive). The sentiment score of each individual emoticon has subsequently been determined as the score closest to the average of the annotators' scores for that emoticon, thus assigning equal weights to the annotators' opinions. For about 88% of our emoticons, the three annotators assigned identical scores to the respective emoticons. The lexicon thus generated is utilized in two implementations of our framework, i.e., one for polarity classification of Dutch documents, and the other for polarity classification of English documents.

3.4.1.4 Evaluation

The implementation of our proposed polarity classification framework allows us to perform experiments in order to compare the performance of several configurations of our framework. First, we consider an absolute baseline of not accounting for the information conveyed by emoticons (Baseline), thus essentially reducing our analysis to an existing lexicon-based document-level polarity classification method (Bal et al., 2011). Then, as a first alternative, we consider an approach in which the sentiment conveyed by emoticons is assumed to affect the surrounding text on a sentence level (Emo.S). Last, we consider accounting for the sentiment conveyed by emoticons on a paragraph level when classifying the polarity of a piece of text (Emo.P).

We assess the performance of each of our considered methods in terms of precision, recall, and F_1-score for positive and negative documents separately, as well as the overall accuracy and macro-level F_1-score. Precision is the proportion of the positively (negatively) classified documents which have an actual classification of positive (negative), whereas recall is the proportion of the actual positive (negative) documents which are also classified as such. The F_1-score is the harmonic mean of precision and recall. The macro-level F_1-score is the average of the F_1-scores of the positive and negative documents, weighted for their respective relative frequencies. Accuracy is the proportion of correctly classified documents.

In order to get a clear view on the impact of accounting for the sentiment conveyed by emoticons in sentiment analysis, we assess the statistical significance of the observed performance differences by means of a paired two-sample two-tailed t-test. To this end, we randomly split our data sets into ten equally sized subsets, on which we assess the performance of our considered methods. The mean performance measures over these subsets can then be compared by means of the paired t-test.

3.4.2 Experimental Results on Dutch Tweets and Forum Posts

On our collection of Dutch tweets and forum posts, our considered polarity classification approaches exhibit clear differences in terms of performance, as demonstrated in Tables 3.3, 3.4, and 3.5. The absolute baseline of not accounting for the sentiment conveyed by emoticons (Baseline) is outperformed by both considered methods of harvesting information from emoticons for the sentiment analysis process. Overall, sentence-level accounting for emoticon sentiment (Emo.S) yields a statistically significant increase in accuracy and macro-level F_1 from 22% to 59% and from 22% to 65%, respectively. Assuming the sentiment conveyed by emoticons to affect the surrounding text on a paragraph level (Emo.P) significantly increases both overall polarity classification accuracy and macro-level F_1 even further to 94%.

The performance differences between our novel Emo.S and Emo.P methods suggest that, in our Dutch corpus, the scope of influence of the sentiment conveyed by emoticons is not always limited to the surrounding text on a sentence level, but may extend to the surrounding text on a paragraph level as well. In general, assuming a paragraph-level influence of emoticons yields significantly better polarity classification performance than assuming a sentence-level influence. Nevertheless, both approaches work significantly better than the Baseline approach, which assumes no influence of emoticons at all.

The comparably weak performance of our Baseline method suggests that in our Dutch documents, emoticons do not often emphasize sentiment that is already conveyed by sentiment-carrying words. Conversely, the authors of our considered tweets and forum posts mostly use emoticons to express or disambiguate their sentiment. This holds for the positive documents, as well as for the negative documents. As such, in our Dutch texts, emoticons are crucial proxies for people's sentiment, as they often capture sentiment that cannot typically be inferred from the sentiment-carrying words used in our texts. This confirms that accounting for the sentiment conveyed by emoticons is a viable strategy when performing sentiment analysis of text.

3.4.3 Validation on English App Reviews

In order to validate our findings presented in Section 3.4.2, we have assessed the performance of our considered polarity classification approaches on a collection of documents in another language, covering another domain. Tables 3.6, 3.7, and 3.8 present the performance of our methods on this collection of English app reviews, some of which contain emoticons.

	Positive			Negative			Overall	
Method	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy	F_1
Baseline	0.222	0.209	0.215	0.216	0.229	0.222	0.219	0.219
Emo.S	0.670	0.650	0.660	0.680	0.590	0.632	0.590	0.646
Emo.P	**0.935**	**0.954**	**0.944**	**0.951**	**0.930**	**0.940**	**0.942**	**0.942**

Table 3.3: The performance of our considered methods on our collection of Dutch tweets and forum posts. The best performance is printed in bold for each performance measure.

Benchmark	Baseline	Emo.S	Emo.P
Baseline	0.000	1.697***	3.308***
Emo.S	-0.629***	0.000	0.597***
Emo.P	-0.768***	-0.374***	0.000

Table 3.4: Relative differences of the overall accuracy of our methods, benchmarked against one another on our collection of Dutch tweets and forum posts. Performance differences marked with * are statistically significant at $p < 0.05$, those marked with ** are significant at $p < 0.01$, and those marked with *** are significant at $p < 0.001$.

Benchmark	Baseline	Emo.S	Emo.P
Baseline	0.000	1.955***	3.310***
Emo.S	-0.662***	0.000	0.458***
Emo.P	-0.768***	-0.314***	0.000

Table 3.5: Relative differences of the macro-level F_1-score of our methods, benchmarked against one another on our collection of Dutch tweets and forum posts. Performance differences marked with * are statistically significant at $p < 0.05$, those marked with ** are significant at $p < 0.01$, and those marked with *** are significant at $p < 0.001$.

Accounting for emoticons in the polarity classification process has a small, yet significant effect on the polarity classification performance when considering all 10,069 documents in this corpus. The overall accuracy significantly increases with about 1% from 74% for the Baseline approach to 75% for both considered emoticon-guided polarity classification methods. Similarly, the macro-level F_1-score exhibits a significant increase of approximately 1% from 73% for the Baseline method to 74% for our novel Emo.S and Emo.P approaches. The differences between the latter emoticon-guided polarity classification approaches are very small and statistically insignificant on the full data set.

The observed performance improvements of our emoticon-guided approaches, compared to the Baseline method, are mainly driven by improved polarity classification performance on the negative documents in our corpus. The Baseline method's performance may be thwarted by the reported tendency of people to write negative texts with rather positive words (Chenlo et al., 2013; Heerschop et al., 2011a; Taboada et al., 2008).

	Positive			Negative			Overall	
Method	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy	F_1
Baseline	0.788	0.856	0.821	0.587	0.469	0.521	0.739	0.730
Emo.S	0.793	0.859	0.825	0.600	0.484	0.536	0.746	0.737
Emo.P	**0.794**	**0.860**	**0.826**	**0.602**	**0.486**	**0.538**	**0.747**	**0.738**

Table 3.6: The performance of our considered methods on our collection of English app reviews. The best performance is printed in bold for each performance measure.

Benchmark	Baseline	Emo.S	Emo.P
Baseline	0.000	0.009***	0.010***
Emo.S	-0.009***	0.000	0.001
Emo.P	-0.010***	-0.001	0.000

Table 3.7: Relative differences of the overall accuracy of our methods, benchmarked against one another on our collection of English app reviews. Performance differences marked with * are statistically significant at $p < 0.05$, those marked with ** are significant at $p < 0.01$, and those marked with *** are significant at $p < 0.001$.

Benchmark	Baseline	Emo.S	Emo.P
Baseline	0.000	0.010***	0.011***
Emo.S	-0.010***	0.000	0.001
Emo.P	-0.011***	-0.001	0.000

Table 3.8: Relative differences of the macro-level F_1-score of our methods, benchmarked against one another on our collection of English app reviews. Performance differences marked with * are statistically significant at $p < 0.05$, those marked with ** are significant at $p < 0.01$, and those marked with *** are significant at $p < 0.001$.

Conversely, the Emo.S and Emo.P methods compensate for this possible bias towards positivity by harvesting crucial information from emoticons. These results support our underlying assumption of the sentiment conveyed by nonverbal cues (i.e., emoticons) dominating the sentiment conveyed by verbal cues – especially for the negative app reviews in our corpus, emoticons appear to play a crucial role in expressing or disambiguating an author's sentiment.

Only about 7% of the documents in our collection of English app reviews contain emoticons. Therefore, the (significant) differences in terms of polarity classification performance of our considered methods are rather small on the full data set. An additional assessment of the performance of our methods on all English app reviews that contain emoticons provides more insight into the potential of our methods. Tables 3.9, 3.10, and 3.11 demonstrate more apparent differences in performance on the 655 app reviews that contain emoticons.

	Positive			Negative			Overall	
Method	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy	F_1
Baseline	0.867	0.863	0.865	0.446	0.453	0.450	0.783	0.784
Emo.S	0.952	0.903	0.927	0.671	0.813	0.735	0.885	0.889
Emo.P	**0.964**	**0.911**	**0.937**	**0.701**	**0.859**	**0.772**	**0.901**	**0.904**

Table 3.9: The performance of our considered methods on English app reviews that contain emoticons. The best performance is printed in bold for each performance measure.

Benchmark	Baseline	Emo.S	Emo.P
Baseline	0.000	0.131***	0.150***
Emo.S	-0.116***	0.000	0.017
Emo.P	-0.131***	-0.017	0.000

Table 3.10: Relative differences of the overall accuracy of our methods, benchmarked against one another on English app reviews that contain emoticons. Performance differences marked with * are statistically significant at $p < 0.05$, those marked with ** are significant at $p < 0.01$, and those marked with *** are significant at $p < 0.001$.

Benchmark	Baseline	Emo.S	Emo.P
Baseline	0.000	0.135***	0.154***
Emo.S	-0.119***	0.000	0.017
Emo.P	-0.133***	-0.017	0.000

Table 3.11: Relative differences of the macro-level F_1-score of our methods, benchmarked against one another on English app reviews that contain emoticons. Performance differences marked with * are statistically significant at $p < 0.05$, those marked with ** are significant at $p < 0.01$, and those marked with *** are significant at $p < 0.001$.

The performance of our methods on the English app reviews that contain emoticons exhibits a pattern that is similar to our findings on our collection of Dutch tweets and forum posts containing emoticons, as presented in Section 3.4.2. Our novel polarity classification method that accounts for emoticons on a sentence level (Emo.S) significantly outperforms the Baseline method with about 13% (in relative terms), with an overall accuracy and a macro-level F_1-score increasing from about 78% to approximately 89%. Accounting for emoticons on a paragraph level (Emo.P) yields a further relative increase of the overall polarity classification performance with about 2%, with an overall accuracy and a macro-level F_1-score amounting to over 90%. Compared to our findings on our Dutch corpus, this improvement's p-value of 0.072 provides a weaker indication that assuming a paragraph-level influence of emoticons is to be preferred over assuming a sentence-level influence. Nevertheless, both emoticon-guided polarity classification approaches significantly outperform our baseline of not accounting for emoticons at all.

Interestingly, in contrast with its observed weak performance on our collection of Dutch tweets and forum posts containing emoticons, the Baseline method performs rather well on our English app reviews that contain emoticons. This suggests that in most of the latter documents, emoticons convey sentiment that is conveyed by the sentiment-carrying words in these documents as well. As such, emoticons form an equally good proxy for an author's sentiment as the sentiment-carrying words in such app reviews. Nevertheless, emoticons play a crucial role in a subset of our English app reviews containing emoticons, where the main purpose of these emoticons is to express or disambiguate an author's intended sentiment. This mainly holds for most of the negative reviews, where the authors have a tendency of using rather positive sentiment-carrying words and negative emoticons in order to convey their negative sentiment. Properly accounting for the sentiment conveyed by emoticons in such cases yields a significantly improved overall polarity classification performance, thus confirming that our proposed method of accounting for the sentiment conveyed by emoticons is not only a viable strategy in our initial corpus of Dutch tweets and forum posts, but in our validation corpus as well.

3.4.4 Caveats

Experiments in recent competitions for sentiment analysis, such as the SemEval 2007 Task 14 on Affective Text (Strappavara and Mihalcea, 2007), have shown how difficult it is to extract the valence (sentiment) of text for both supervised and unsupervised approaches, which currently lag behind the performance of the inter-annotator agreement for valence. In this light, our results clearly indicate that considering emoticons when analyzing sentiment on natural language text appears to be a fruitful addition to the state-of-the-art of (lexicon-based) sentiment analysis. Our results suggest that whenever emoticons are used, these visual cues play an important, if not crucial role in conveying an author's sentiment. The sentiment conveyed by emoticons tends to dominate the sentiment conveyed by verbal cues in both of our considered corpora. As such, emoticons have proven to be helpful indicators of intended sentiment.

However, some issues still remain to be solved. One source of polarity classification errors lies in the interpretation of a text by human readers and the preference of these readers for certain aspects of the text over others. Consider, e.g., the fragment "*Interesting product =D Just not for me... =/*". Our framework would assign a sentiment score of 0 to this text, as the emoticons cancel each other out in this particular text. However, in the annotation process of our set of Dutch tweets and forum messages, our three human annotators initially did not typically agree on the overall polarity of fragments like this.

All three of our human annotators turned out to deem one part of such a fragment to be more important for conveying the overall sentiment than the other part, even though they initially did not agree on which part was crucial for the polarity of the fragment. Conversely, for our framework, each part of a text contributes equally to conveying the overall sentiment of the text.

Another source of errors can be nicely illustrated when analyzing the polarity conveyed by our English app reviews. The reviews in our corpus often start with a description of the app. These descriptions may already contain sentiment-carrying words, whereas the writer is not yet expressing his or her own opinion at that stage of the review. Apparently, aspects other than sentiment-carrying words and emoticons, such as their positioning or rhetorical role (Heerschop et al., 2011a), may be worthwhile exploiting in sentiment analysis.

3.5 Conclusions

With people increasingly using emoticons in their virtual utterances of opinions, it is of paramount importance for automated polarity classification tools to correctly interpret these graphical cues for sentiment and to account for their effects on other cues on a semantic level. Our key contribution lies in our analysis of the role that emoticons typically play in conveying a text's overall sentiment, as well as in the proposal and evaluation of our novel method for exploiting emoticons in lexicon-based polarity classification.

Whereas emoticons have until now been considered to be used in a way similar to how textual cues for sentiment are used (Thelwall et al., 2010), the qualitative analysis presented in this chapter demonstrates that the sentiment associated with emoticons typically dominates the sentiment conveyed by textual cues in a text segment. The results of our analysis indicate that people typically use emoticons in natural language text in order to express, emphasize, or disambiguate their sentiment in particular text segments, thus rendering them potentially better local proxies for people's intended overall sentiment than textual cues.

On a collection of Dutch tweets and forum messages, as well as on another collection of English app reviews, we find that accounting for the sentiment conveyed by emoticons on a paragraph level – and, to a lesser extent, on a sentence level – significantly improves the performance of a lexicon-based polarity classifier. Our findings suggest that whenever emoticons are used, their associated sentiment dominates the sentiment conveyed by textual cues and forms a good proxy for the polarity of natural language text.

We have demonstrated how to jointly use explicit textual cues and emoticons when classifying the polarity of a text, by incorporating linguistic analysis on a semantic level into the polarity classification process. A possible direction for future work could be to further explore and exploit the interplay of emoticons and textual cues for sentiment, for instance in cases when emoticons are used to intensify sentiment that is already conveyed by the text, or when sarcastic emoticons are used that may invert rather than override the polarity conveyed by the text. Another interesting direction for future work would be to further incorporate semantic analysis into the polarity analysis process. This is done in the next chapter of this dissertation.

An interesting finding of our work is that our human annotators considered some parts of texts to be more important than other parts of these texts, in terms of their relevance for the texts' conveyed overall sentiment. This perceived importance appears to be related to structural aspects of content. Therefore, we envisage a fruitful direction for future research to lie not so much in further exploitation of explicit cues like words and emoticons, but rather in the exploitation of latent cues like the positioning or role of text segments. The latter cues could aid the differentiation between important and less important text segments for polarity classification in future work. This idea is further explored in Chapters 5, 6, and 7.

Chapter 4

Exploiting Semantics for Sentiment Analysis in a Multi-lingual Setting*

MANY *sentiment analysis methods rely on sentiment lexicons, containing words and their associated sentiment, and are tailored to one specific language. Yet, the ever-growing amount of data in different languages on the Web renders multilingual support increasingly important. We assess various methods for supporting an additional target language in lexicon-based sentiment analysis. As a baseline, we automatically translate text into a reference language for which a semantically enabled sentiment lexicon is available. Second, we consider mapping sentiment scores from this reference sentiment lexicon to a new sentiment lexicon for the target language, by traversing relations between language-specific semantic lexical resources. Last, we consider creating a new target sentiment lexicon by propagating sentiment of seed words in a semantic lexicon for the target language. When extending sentiment analysis from English to Dutch, mapping sentiment across languages by exploiting relations between semantic lexicons yields a significant performance improvement over machine translation on our data. Propagating sentiment in language-specific semantic lexicons can outperform our baseline even further, depending on the seed set of sentiment-carrying words. This indicates that sentiment is linked to semantics and, moreover, tends to be somewhat language-specific as well.*

*This chapter is based on the article "A. Hogenboom, B. Heerschop, F. Frasincar, U. Kaymak, and F. de Jong. Multi-lingual Support for Lexicon-Based Sentiment Analysis Guided by Semantics. *Decision Support Systems*, 66(1):43–53, 2014."

4.1 Introduction

In today's complex, globalizing markets, information monitoring tools are of paramount importance for decision makers. Such tools help decision makers in identifying issues and patterns that matter, as well as in tracking and predicting emerging events. Recent advances in tools for information monitoring and extraction have been primarily focused on retrieving explicit pieces of information from natural language text on different levels of granularity (Chang et al., 2006).

The state-of-the-art of information monitoring and extraction tools enables us to, e.g., identify entities like companies, products, or brands in text, and to subsequently extract more complex concepts such as events in which these entities play various roles (Hogenboom et al., 2013c). Recent research endeavors additionally explore how to perform such information extraction tasks on a multitude of heterogeneous sources in an ever-changing environment (Chan, 2006; Chang et al., 2003; Tari et al., 2012).

However, latent pieces of information can be extracted from natural language text as well. For instance, recent work has made it possible to detect the distinct topics that people discuss in their (on-line) conversations (Cui et al., 2011; Wang et al., 2012). Yet, it is not so much the entities, events, or topics that people discuss per se, but rather people's sentiment with respect to these subjects that provides decision makers with valuable information. For example, consumer sentiment has been shown to have a significant impact on stock ratings (Schumaker et al., 2012; Yu et al., 2013) and sales (Ghose and Ipeirotis, 2011; Yu et al., 2012). Additionally, the importance of tracking one's stakeholders' sentiment has been demonstrated for economic systems (Ludvigson, 2004), financial markets (Arnold and Vrugt, 2008), politics (Baron, 2005), organizations (Holton, 2009), and reputation management (Jansen et al., 2009).

Recent developments on the Web enable users to produce an ever-growing amount of virtual utterances of opinions or sentiment through, e.g., messages on Twitter, blogs, or reviews, in any language of their preference. The analysis of sentiment in the overwhelming amount of available multi-lingual textual data is challenging at best. This challenge can be addressed by means of automated sentiment analysis techniques, focusing on determining the polarity of natural language text. Typical approaches involve scanning a text for cues signalling its polarity, e.g., (parts of) words or other (latent) features of natural language text. Lexicon-based sentiment analysis approaches have gained (renewed) attention in recent work (Cesarano et al., 2006; Devitt and Ahmad, 2007; Ding et al., 2008; Heerschop et al., 2011a,c; Hogenboom et al., 2012b; Taboada et al., 2011), not in the least due to their comparably robust performance across domains and texts (Taboada et al., 2008).

Such methods essentially rely on lexical resources containing words and their associated sentiment, i.e., sentiment lexicons, and their nature allows for intuitive ways of accounting for structural or semantic aspects of text in sentiment analysis (Heerschop et al., 2011a; Hogenboom et al., 2010b).

Many existing lexicon-based sentiment analysis approaches are tailored to one specific language – typically English. However, in order for automated sentiment analysis to be useful for decision makers in today's complex, globalizing markets, automated sentiment analysis tools need to be able to support multiple languages rather than English only. Therefore, we explore how we can analyze sentiment in another language – i.e., Dutch – for which we have nothing more but some morphologic, lexical, and syntactic parsing tools, a semantic lexical resource, and a handful of positive and negative sample words.

A good starting point is SentiWordNet (Baccianella et al., 2010; Esuli and Sebastiani, 2006), as recent research has proven this large semantically enabled sentiment lexicon for English, generated by means of machine learning techniques, to be rather effective when used for analyzing sentiment in texts published in our reference language, i.e., English (Heerschop et al., 2011b). As a first step, one could consider translating texts from a target language, i.e., Dutch, to our reference language, i.e., English, in order to be able to subsequently utilize the well-established SentiWordNet sentiment lexicon for the reference language in the sentiment analysis process.

However, as subjectivity is associated with word meanings rather than words (Mihalcea et al., 2007), literal translation of texts to a reference language in order to benefit from the available sentiment lexicon for the reference language may be suboptimal in automated sentiment analysis of texts in another language. As an alternative, we therefore propose to map the sentiment from the reference sentiment lexicon to a sentiment lexicon for the target language, by means of traversing relations between large language-specific semantic lexical resources, thus accounting for word meanings rather than lexical representations. Additionally, we consider an approach that involves propagating sentiment from a seed set of words in a language-specific semantic lexical resource for each considered language separately, in order to generate language-specific sentiment lexicons which can subsequently be used in language-specific sentiment analysis methods.

The remainder of this chapter is organized as follows. In Section 4.2, we discuss related work on (multi-lingual) sentiment analysis and how semantics may be exploited in this process. We then elaborate on our framework for assessing our considered methods for dealing with another language in sentiment analysis in Section 4.3. Our findings are discussed in Section 4.4. We conclude and provide directions for future work in Section 4.5.

4.2 Related Work

Many methods have already been proposed in order to deal with language-specific sentiment analysis problems (Cambria et al., 2013; Feldman, 2013; Liu, 2012; Pang and Lee, 2008), i.e., methods focused on extracting subjective information such as polarity from natural language text written in specific languages. However, the exploration of how to support multiple languages when analyzing sentiment has only just begun.

4.2.1 Multi-lingual Sentiment Analysis

Today's sentiment analysis systems must be able to deal with an abundance of multilingual sentiment-carrying user-generated content. As different approaches are required for distinct languages (Boiy and Moens, 2009), existing work does not typically focus on devising a single sentiment analysis approach for multiple languages, but rather on analyzing the sentiment conveyed by documents in selected languages, mainly by means of applying sentiment analysis techniques tailored to each specific language. Existing work is primarily focused on devising sentiment analysis methods for other languages with minimal effort, without sacrificing too much accuracy. Rather than constructing new frameworks for languages other than the reference language (Abbasi et al., 2008; Boiy and Moens, 2009; Dai et al., 2007a,b; Gliozzo and Strapparava, 2005), recent work uses machine translation techniques in order to be able to re-use many existing tools when performing automated sentiment analysis on multi-lingual textual content.

Sentiment analysis of machine-translated texts may seem a rather ineffective approach, as machine translation typically fails to correctly translate substantial amounts of text and moreover tends to reduce well-formed texts to sentence fragments. Nevertheless, recent work on sentiment analysis of news messages in nine languages demonstrates that sentiment classification accuracy is largely independent of the quality of the machine translator used (i.e., the translator does not necessarily have to produce well-formed texts) and that sentiment analysis of texts that have been translated into English is consistent across languages, after normalizing sentiment scores in order to allow for meaningful cross-cultural comparisons (Bautin et al., 2008).

Other existing work suggests that in some cases, sentiment analysis of machine-translated texts can yield even better results than sentiment analysis of the original texts, especially when the original language is not easily interpreted by state-of-the-art natural language processing tools. For instance, Wan (2008) uses a Chinese sentiment analysis framework for classifying the sentiment of Chinese reviews, and an English framework for classifying the sentiment of these Chinese reviews after machine translation into English.

Wan (2008) shows that sentiment analysis of the translated texts outperforms sentiment analysis of the original texts. An ensemble of both methods further improves performance.

Machine translation can be utilized in another way as well in order to facilitate automated sentiment analysis in multiple languages. Rather than performing sentiment analysis on machine-translated texts, many researchers focus on automatically generating sentiment lexicons by using machine translation. A common approach is to automatically translate an existing sentiment lexicon (Mihalcea et al., 2007), and, possibly, to subsequently propagate the sentiment to semantically related words (Jijkoun and Hofmann, 2009). An alternative approach, which has proven to outperform machine translation of sentiment lexicons, is to automatically generate a sentiment lexicon from a collection of (automatically) translated and annotated texts (Banea et al., 2008, 2010; Lin et al., 2012; Mihalcea et al., 2007). However, research suggests that the subjectivity of most of the words in sentiment lexicons is lost in translation – subjectivity appears to be a property associated not with words, but with word meanings (Mihalcea et al., 2007). Semantic lexicons can be used in order to address this issue.

4.2.2 Semantic Lexicons

A widely used on-line semantic lexical resource is WordNet (Fellbaum, 1998), the design of which has been inspired by psycholinguistic theories of human lexical memory. WordNet is organized into sets of cognitive synonyms – synsets – which can be differentiated based on their Part-of-Speech (POS) type. Each WordNet synset expresses a distinct concept and is linked to other synsets through different kinds of relations, such as synonymy, antonymy, hyponymy, or meronymy. The need for such a lexical reference system has arisen as conventional dictionaries do not usually capture such semantic relations. Conventional dictionaries use lexicographical sorting for words for human users' convenience. Conversely, WordNet has been designed to be used under program control and enables the distinction between different word forms and word meanings.

SentiWordNet (Baccianella et al., 2010; Esuli and Sebastiani, 2006) is a lexical resource in which each WordNet synset is associated with three numerical scores, quantifying its associated sentiment. These scores describe how objective, positive, and negative the terms contained in a synset are. An ensemble of eight ternary classifiers has been used to classify each synset as either objective, positive, or negative, based on a vector representation of the associated description of the synset. The overall objectivity, positivity, and negativity scores for a synset have then been determined as the (normalized) proportions of the classifiers that assigned the corresponding labels to the synset.

The availability of semantic lexical resources is not limited to the English language. For instance, EuroWordNet (Vossen, 1997) has been developed as a collection of semantic lexicons for several European languages, including English, Dutch, Italian, and Spanish. For each supported language, a semantic lexicon has been created, with a structure similar to WordNet's structure. Additionally, the language-specific semantic lexicons in EuroWordNet are linked to one another through WordNet, such that each English synset is associated with its equivalents in the languages covered by EuroWordNet.

For Dutch, i.e., the language we consider as an alternative to our English reference language, a more extensive semantic lexicon has been developed on top of EuroWordNet as well. In DutchWordNet (Cornetto) (Vossen et al., 2007), the Dutch part of EuroWordNet has been enriched with information from the Referentie Bestand Nederland (RBN), which is a lexical database for Dutch, containing information on orthography, morphology, syntax, semantics, pragmatics, and combinatorics.

Language-specific semantic lexical resources and their interlinkage through semantic lexical resources such as WordNet can facilitate new approaches for extending an existing lexicon-based sentiment analysis approach from one language to another. The semantic relations between language-specific semantic lexicons could be exploited in order to propagate a sentiment lexicon from one language to another, while preserving semantics. Alternatively, sentiment scores for a seed set of words could be propagated through a language-specific semantic lexicon in order to generate language-specific sentiment lexicons (Heerschop et al., 2011b; Hu and Liu, 2004; Kamps et al., 2004; Lerman et al., 2009). As both types of approaches account for semantics, they may compensate for the drawbacks of existing machine translation methods for multi-lingual sentiment analysis.

4.3 Framework

In order to investigate how lexicon-based sentiment analysis can be extended from our reference language, i.e., English, to another language, i.e., Dutch, we first need a lexicon-based sentiment analysis framework for the reference language. This framework can then serve as a starting point for an extension to another language.

4.3.1 Polarity Classification

We use a binary polarity classifier (Heerschop et al., 2011b) that classifies documents as either positive or negative based on the aggregated sentiment scores for individual words, which have been retrieved from a semantic sentiment lexicon such as SentiWordNet.

For an arbitrary synset, we compute a single sentiment score based on its objectivity, positivity, and negativity scores (all positive real numbers which sum to 1), by subtracting the negativity score from the positivity score, thus obtaining a real number in the interval $[-1, 1]$, representing sentiment scores in the range from negative to positive, respectively.

In our polarity classification process, detailed in Algorithm 4.1, documents are first split into sentences and words. Then, each word's POS type, lemma, and word sense are determined in order to subsequently retrieve its sentiment score from the sentiment lexicon. For the word sense disambiguation process, we use a Lesk-based algorithm for WordNet (Dao and Simpson, 2005), as described by Heerschop et al. (2011b). The algorithm iteratively selects the word sense that is semantically most similar to the words in the context, i.e., the other words in a sentence.

After retrieving all word-level sentiment scores from the sentiment lexicon, the sentiment score ζ_d of a document d is computed by summing the sentiment scores ζ_t of each non-stopword t in each sentence s of the document, i.e.,

$$\zeta_d = \sum_{s \in d} \sum_{t \in s} \zeta_t. \tag{4.1}$$

The resulting document-level sentiment score ζ_d is subsequently used in order to classify the document's polarity class c_d as either positive (1) or negative (-1), i.e.,

$$c_d = \begin{cases} -1 & \text{if } (\zeta_d - \epsilon) < 0, \\ 1 & \text{otherwise,} \end{cases} \tag{4.2}$$

with ϵ representing an offset correcting a possible bias in sentiment scores caused by people's tendency to write negative texts with rather positive words (Taboada et al., 2008). Following Taboada et al. (2008), we calculate this ϵ on a training set as

$$\epsilon = 0.5 \left(\frac{\sum_{d \in P} \zeta_d}{|\Phi|} + \frac{\sum_{d \in N} \zeta_d}{|N|} \right), \tag{4.3}$$

with Φ denoting the subset of positive documents in the training set, and N denoting the subset of negative documents in the training set.

Our sentiment analysis framework has been developed for classifying the polarity of English documents. As such, in order to be able to classify the polarity of documents written in another language, the latter documents could be automatically translated into the reference language, such that they can be analyzed by means of the sentiment analysis framework for the reference language. Thus, our existing sentiment analysis framework for English documents can be used for classifying the polarity of Dutch documents.

Algorithm 4.1: Classifying a document's polarity.

```
input  : A document $d$ and an offset $\epsilon$
output : The polarity classification $c_d$ of document $d$
$\zeta_d = 0$;
foreach $s \in d$ do
    foreach $t \in s$ do
        $pos$ = findPOS($t$, $s$);
        $lemma$ = findLemma($t$, $pos$);
        $sense$ = findWordSense($t$, $s$, $pos$);
        $\zeta_t$ = getWordScore($lemma$, $sense$, $pos$);
        $\zeta_d = \zeta_d + \zeta_t$;
    end
end
$c_d = 1$;
if $(\zeta_d - \epsilon) < 0$ then
    $c_d = -1$;
end
return $c_d$;
```

However, the concepts behind our English polarity classification framework per se can be applied to polarity classification in Dutch as well, provided that the lexical and syntactical parsing tools for identifying sentences, words, POS, and lemmas are available, as well as a semantic lexical resource for the Dutch language. The latter semantic lexical resource can be used for word sense disambiguation, as well as for constructing a Dutch sentiment lexicon that can be used in a sentiment analysis framework with components tailored to the Dutch language.

Our framework (visualized in Figure 4.1) supports two of such alternatives to the machine translation approach. First, we consider traversing the relations between language-specific semantic lexicons in order to map the existing sentiment lexicon for the English reference language to a new sentiment lexicon for the Dutch target language. This method is detailed in Section 4.3.2. Second, we consider propagating sentiment within language-specific semantic lexical resources, as described in Section 4.3.3.

4.3.2 Traversing Relations between Semantic Lexicons

The valuable information contained in the sentiment lexicon of an existing sentiment analysis approach for the reference language can be utilized in another language when the information is used to generate a sentiment lexicon for the target language. This may be done by (automatically) translating an existing sentiment lexicon from the reference language into the target language (Jijkoun and Hofmann, 2009; Mihalcea et al., 2007).

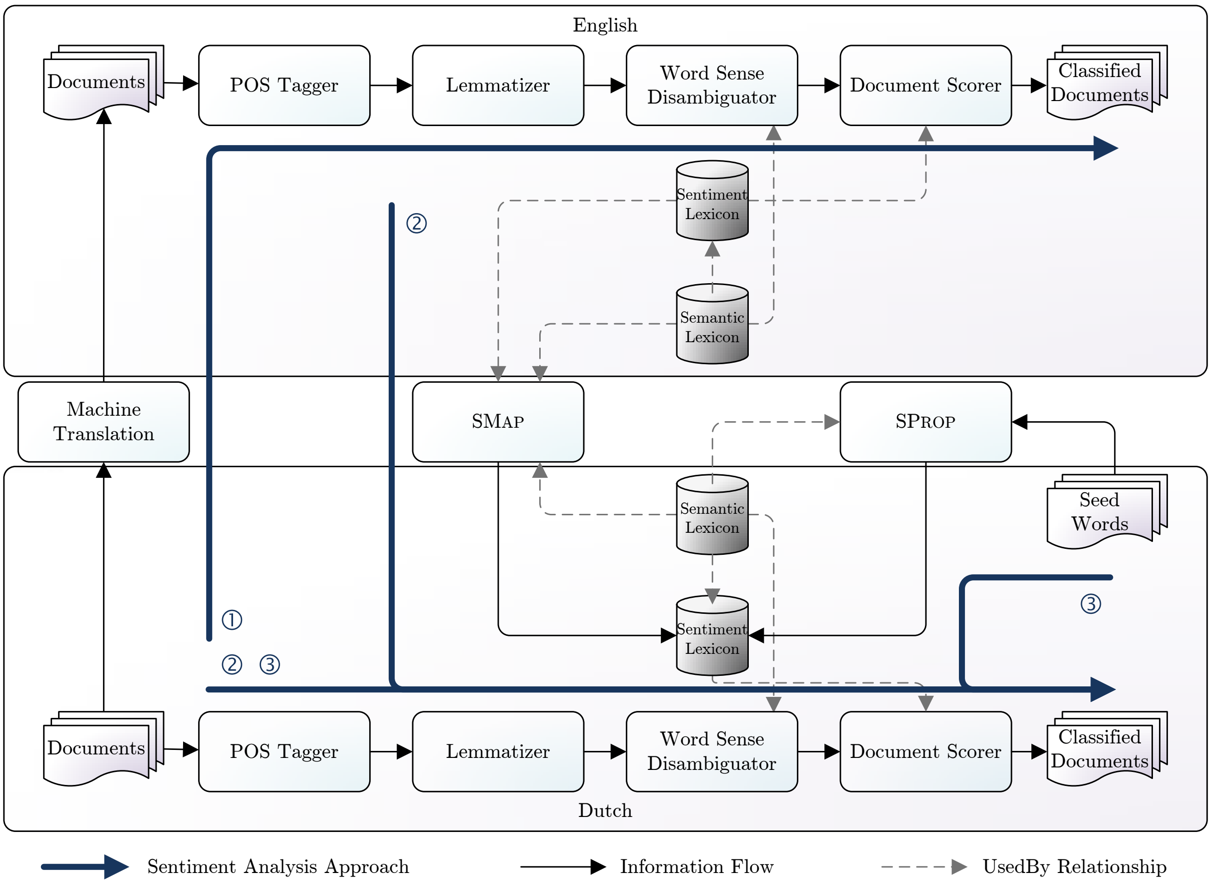


Figure 4.1: Our sentiment analysis framework, with language-specific components for both English and Dutch. Our three considered approaches of using these components in order to analyze the sentiment of Dutch documents are marked with bold arrows. Approach ① is to translate our Dutch documents into English and to subsequently use the available existing English sentiment analysis components. The alternative approaches ② and ③ involve analyzing the sentiment of our Dutch documents by means of Dutch language-specific components while exploiting a sentiment lexicon that has been constructed based on either an existing English sentiment lexicon (②), or seed sets of Dutch sentiment-carrying words (③).

However, subjectivity tends to be associated with word meanings rather than with lexical representations of words alone (Mihalcea et al., 2007). Therefore, we propose a novel method of translating a sentiment lexicon from a reference language to a target language, while taking into account the semantics of the words in the sentiment lexicons. In order to accomplish this, we make use of language-specific semantic lexical resources and their interrelations.

In our proposed cross-lingual sentiment score mapping method SMap (illustrated in Figure 4.2), we assume an existing sentiment lexicon for the reference language to be linked to a semantic lexical resource with meaningfully related words and concepts (synsets).

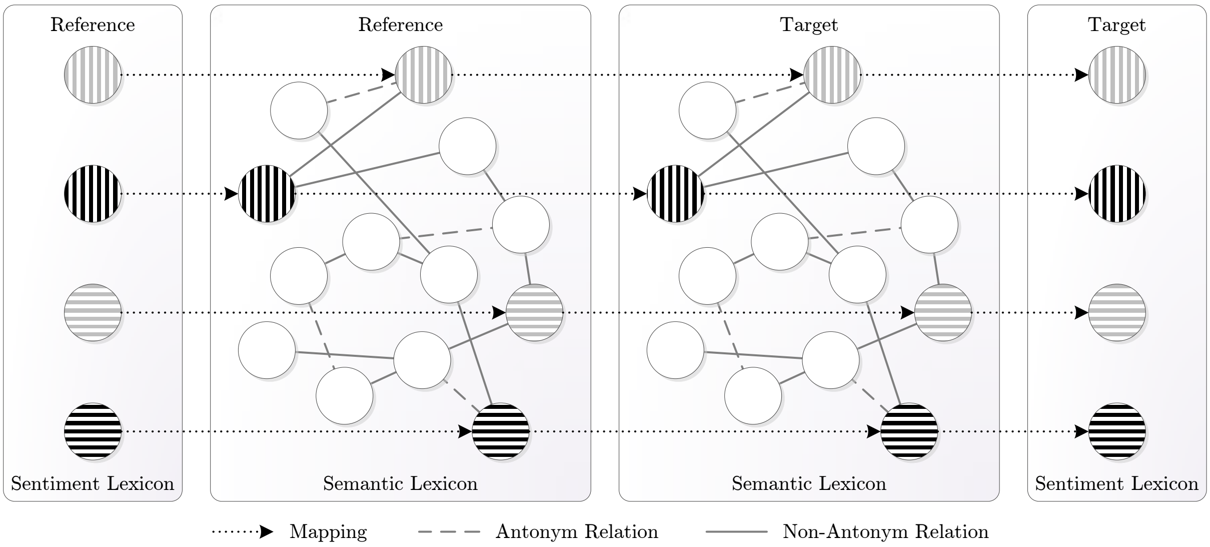


Figure 4.2: Our novel method for mapping sentiment scores from a reference language to a target language. Positive words and synsets are marked with vertical stripes, whereas negative words and synsets are marked with horizontal stripes. Others are left blank. Darker shading implies stronger sentiment.

Provided that a mapping exists between this semantic lexicon and an equivalent semantic lexicon for another language, the sentiment from the reference sentiment lexicon can be mapped to a new sentiment lexicon for the target language by traversing the associated relations between the semantic lexicons of both respective languages.

For example, the English SentiWordNet sentiment lexicon can be used as a starting point for our proposed cross-lingual sentiment score mapping procedure. SentiWordNet contains information on the sentiment associated with all synsets in the WordNet semantic lexicon. Additionally, a mapping exists between WordNet and its Dutch equivalent DutchWordNet (Cornetto). By exploiting these relations, SentiWordNet sentiment scores associated with English WordNet synsets can be projected onto equivalent Dutch synsets in DutchWordNet (Cornetto), thus yielding a Dutch sentiment lexicon.

In order to propagate sentiment associated with synsets through language-specific semantic lexical resources, we first map the synsets in the reference sentiment lexicon to the reference semantic lexicon. Additionally, we map the synsets in the semantic lexicon for the new language to their equivalent synsets in the reference semantic lexicon. Subsequently, for each synset in the reference sentiment lexicon, we use these mappings to find the equivalent synsets and their synonyms in the semantic lexicon for the target language. These synsets are then assigned the sentiment score of the respective synsets in the reference sentiment lexicon. The result is saved in the new sentiment lexicon for the target language. This process is further detailed in Algorithm 4.2.

Algorithm 4.2: Sentiment propagation via relations between semantic lexicons.

```
input  : The reference sentiment lexicon $S^*$, the reference semantic lexicon $L^*$, and the
         semantic lexicon for the target language $L'$
output : The sentiment lexicon for the target language $S'$
$S' = \emptyset$;
foreach $sentiWord^* \in S^*$ do
    $synset^*$ = getSynset($sentiWord$, $S^*$);
    $synset'$ = mapSynsetFromTo($synset^*$, $L^*$, $L'$);
    if $synset' \neq \emptyset$ then
        $pos$ = getPos($synset^*$, $L^*$);
        $\zeta$ = getScore($synset^*$, $S^*$);
        $synonyms$ = getSynonyms($synset'$, $L'$);
        foreach $t \in synonyms$ do
            $lemma$ = getLemma($t$, $L'$);
            $sense$ = getWordSense($t$, $L'$);
            $S' = \{S', \{synset', lemma, sense, pos, \zeta\}\}$;
        end
    end
end
return $S'$;
```

4.3.3 Propagation of Sentiment within Semantic Lexicons

When creating a new sentiment lexicon for a target language, one could also consider not to use a reference sentiment lexicon as a starting point, as the sentiment associated with words or word meanings may have a cultural dimension. Instead, one could consider creating a new sentiment lexicon for the target language by propagating the sentiment of a small seed set of words to words which are semantically related (Heerschop et al., 2011b; Hu and Liu, 2004; Kamps et al., 2004; Lerman et al., 2009).

In our sentiment propagation method SProp (detailed in Algorithms 4.3 and 4.4 and visualized in Figure 4.3), semantic relations in a language-specific semantic lexicon are traversed for each seed word. Examples of such semantic relations are hyponymy (type-of relations), synonymy, and antonymy. In the sentiment propagation process, each encountered word t is stored with a sentiment score ζ_t, based on the score ξ of the seed word, a diminishing factor δ, and the number of steps k (with a maximum of K) between the seed word and t, i.e.,

$$\zeta_t = \xi \tau \delta^k, \quad \tau \in \{-1, 1\}, \quad k \in \{1, \dots, K\}, \quad -1 \leq \xi \leq 1, \quad 0 < \delta < 1, \tag{4.4}$$

with τ indicating whether to invert (-1) the score, i.e., when traversing antonym relations, or not (1).

Algorithm 4.3: Sentiment propagation in a language-specific semantic lexicon.

```
input  : The semantic lexicon for the target language $L'$, a list *seeds* of the words to
         propagate, the maximum number of iterations $K$, and a diminishing factor $\delta$
output : A sentiment lexicon $S'$ containing all propagated words with their computed
         sentiment scores
$syn$ = getSynsets($L'$);
$S' = \emptyset$;
foreach $t \in seeds$ do
    $\xi$ = score($t$);
    $S'$ = propWord($syn$, $S'$, $t$, $\xi$, $\delta$, 1, $K$); // See Algorithm 4.4
end
return $S'$;
```

In each iteration of our algorithm, the sentiment score ζ_t for a word t is propagated to the words in its directly related synsets. With each next traversed semantic relation, the propagated sentiment is further diminished. Thus, words that are semantically more closely related to a seed word obtain a higher (absolute) sentiment score than those with a more indirect semantic relation to a seed word. If a word is encountered multiple times in the process of propagating the sentiment associated with seed words, this word is assigned the score obtained from the shortest path between the word and any of the seeds, as we assume that the shorter the path, the more accurate the sentiment can be determined.

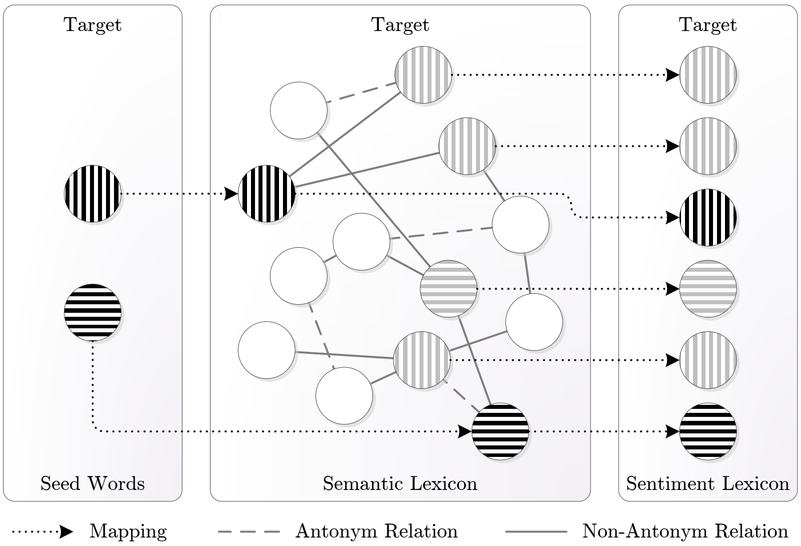


Figure 4.3: Our proposed method for propagating the sentiment of a set of seed words through the semantic lexicon of a target language. The sentiment lexicon thus generated in one propagation step is visualized. Positive words and synsets are marked with vertical stripes, whereas negative words and synsets are marked with horizontal stripes. Others are left blank. Darker shading implies stronger sentiment.

Algorithm 4.4: Propagating a word's sentiment in a lexical resource (propWord).

```
input  : All *synsets* in the semantic lexicon for the target language, a sentiment lexicon
         for the target language $S'$, a word $t$ to propagate, the score $\xi$ of $t$, a
         diminishing factor $\delta$, the current iteration $k$, and the maximum number of
         iterations $K$
output : A sentiment lexicon $S'$ containing all propagated words with their computed
         sentiment scores
if $k \leq K$ then
    $reachedIn$ = getSteps($t$); // $\infty$ for new $t$
    if $reachedIn > k$ then
        $synsetsWithWord$ = getSynsets($synsets$, $t$);
        foreach $synset \in synsetsWithWord$ do
            $pos$ = getPOS($synset$);
            $syns$ = getSynonyms($synset$);
            foreach $syn \in syns$ do
                $lemma$ = getLemma($syn$);
                $sense$ = getWordSense($syn$);
                $S' = \{S', \{lemma, sense, pos, \xi\}\}$;
            end
            $rels$ = getRelations($synset$);
            foreach $r \in rels$ do
                $\tau = 1$;
                if $r == antonym$ then
                    $\tau = -1$;
                end
                $rSyns$ = getSynonyms($r$);
                foreach $rw \in rSyns$ do
                    $\xi' = \xi\tau\delta$;
                    $k' = k + 1$;
                    $\theta = \{synsets, S', rw, \xi', \delta, k', K\}$;
                    $S'$ = propWord($\theta$);
                end
            end
        end
        takeSteps($t$, $k$);
    end
end
return $S'$;
```

4.4 Evaluation

Our proposed methods can be used for exploring how lexicon-based sentiment analysis can best be extended from our reference language, i.e., English, to another language, i.e., Dutch. In this section, we present the set-up and results of our experiments.

4.4.1 Experimental Setup

In our experiments, we focus on a set of 600 positive and 600 negative opinionated Dutch documents on 40 distinct topics, crawled from Dutch review Web sites, forums, and blogs. The classifications have been made by three human annotators, until they reached full consensus. On this corpus, we assess the performance of our considered methods by means of the 10-fold cross-validated overall sentiment classification accuracy and the macro-level F_1-score. We assess the statistical significance of performance differences by means of a paired two-sample two-tailed t-test.

The implementation of our sentiment classification framework has been done in C#.Net. We have built upon our existing work for classifying the sentiment of English documents (Heerschop et al., 2011b), which classifies sentiment as described in Section 4.3.1. We have constructed a similar implementation for sentiment classification of Dutch documents, which extends the English implementation by means of our considered translation and sentiment propagation methods as discussed in Section 4.3.

When classifying the sentiment of English documents, we use regular expressions in order to split the text into words. POS tagging is done with a SharpNLP (SharpNLP, 2006) POS tagger. Lemmatization and word sense disambiguation tasks are performed by means of the C# WordNet.Net (Simpson and Crowe, 2005) API for WordNet. Our sentiment classification approach for English documents relies on a semantic lexicon and a sentiment lexicon. We link word senses to WordNet (Fellbaum, 1998), whereas we retrieve the associated sentiment scores from SentiWordNet 3.0 (Baccianella et al., 2010). On a widely used data set of 1,000 positive and 1,000 negative English movie reviews (Pang and Lee, 2004), our approach has an overall sentiment classification accuracy and macro-level F_1-score of approximately 60% (Heerschop et al., 2011b).

The implementation of our sentiment classification method for Dutch documents is similar to our existing implementation for English documents, even though it utilizes different language-specific components. For POS tagging, our current implementation uses a SharpNLP (SharpNLP, 2006) POS tagger. Lemmatization is performed by the Tadpole (van den Bosch et al., 1997) lemmatizer. Word sense disambiguation is done by applying our own implementation of the Lesk-based algorithm Dao and Simpson (2005) implemented in WordNet.Net (Simpson and Crowe, 2005). Our Dutch sentiment classification approach relies on DutchWordNet (Cornetto) (Vossen et al., 2007), a large semantic lexical resource for Dutch. We use this semantic lexicon for word sense disambiguation as well as for sentiment lexicon creation by means of one of our considered methods other than our machine translation baseline.

We consider three main sentiment analysis approaches. In our considered machine translation (MT) baseline, first, we automatically translate the Dutch texts from our considered corpus into English by using the Google Translate service (Google, 2007). Then, we classify the sentiment conveyed by the translated documents by means of our sentiment classification approach for English documents.

Our first alternative to this machine translation baseline is our cross-lingual sentiment score mapping method (SMap), in which we first map the sentiment associated with all WordNet synsets from SentiWordNet 3.0 to all equivalent synsets in DutchWordNet (Cornetto). We subsequently classify the sentiment conveyed by the Dutch documents in our corpus by means of our sentiment classification approach for Dutch text, while utilizing the Dutch sentiment lexicon thus constructed.

As a second alternative, we use the SProp method to propagate the sentiment of a set of seed words through DutchWordNet (Cornetto) and we subsequently classify the conveyed sentiment by using the constructed sentiment lexicon in our sentiment classification method for Dutch documents. We assess the performance of the SProp method when using three different seed sets of positive words (with a sentiment score of 1) and negative words (with a sentiment score of -1). For any of these seed sets, sentiment scores are propagated by traversing the holonym, hyperonym, and hyponym relations between synsets in DutchWordNet (Cornetto), with a maximum number of iterations K of 8 and a diminishing factor δ of 0.9, as an initial exploration of the parameter space indicated that these settings were most promising.

Each of our seed sets of sentiment-carrying words, detailed in Table 4.1, has been manually constructed by three human annotators, all of whom are native Dutch speakers. Our annotators have constructed the Dutch seed sets by combining their knowledge of the Dutch language with the most positive and negative SentiWordNet synsets. The first set contains ten positive and ten negative Dutch words. The second set is an expansion of the first set, containing 26 positive and 17 negative Dutch words. Another expansion has resulted in a third seed set, consisting of 26 positive and 24 negative Dutch words.

4.4.2 Experimental Results

The performance of our considered methods of classifying sentiment conveyed by Dutch documents by exploiting an existing method for sentiment classification of English documents is summarized in Tables 4.2, 4.3, and 4.4. These results demonstrate that some of our approaches work better than others for performing sentiment analysis of documents in another language than the reference language. Several observations can be made.

Initial set		First expansion		Second expansion	
Seed word	ξ	Seed word	ξ	Seed word	ξ
Mooi	1	Super	1	Treurig	-1
Schoon	1	Schitterend	1	Onheilspellend	-1
Aanbiddelijk	1	Hart	1	Griezelig	-1
Duidelijk	1	Amicaal	1	Schelden	-1
Elegant	1	Gezelligheid	1	Irriteren	-1
Beter	1	Goed	1	Vervelen	-1
Glimmend	1	Aanbidden	1	Negatief	-1
Perfect	1	Plezier	1		
Energiek	1	Aangenaam	1		
Trots	1	Uitmuntend	1		
Klote	-1	Beeldig	1		
Boos	-1	Positief	1		
Arrogant	-1	Veilig	1		
Bewolkt	-1	Vrijheid	1		
Verstoord	-1	Vakantie	1		
Onmogelijk	-1	Ontspanning	1		
Haat	-1	Mongool	-1		
Twijfelen	-1	Tering	-1		
Verafschuwen	-1	Wantrouwig	-1		
Imbeciel	-1	Verward	-1		
		Gedachteloos	-1		
		Berucht	-1		
		Jammer	-1		

Table 4.1: Considered sentiment-carrying seed words, with their associated sentiment scores ξ. Our first set contains the words in the initial set of words only. Our second set contains the words in the initial set, as well as the words in the first expansion. Our third set contains the words in the initial set, the first expansion, and the second expansion.

In general, all considered approaches exhibit a rather balanced performance, as they seem to perform equally well when classifying the sentiment of positive and negative documents. Additionally, when exploiting our existing sentiment analysis framework for English texts by means of our considered approaches, the best achievable performance of our framework on Dutch documents is rather comparable to the performance of the existing framework on English documents.

As reported by Heerschop et al. (2011b), our utilized existing English sentiment analysis approach can obtain an overall accuracy and macro-level F_1-score of up to about 60% on a widely used collection of English movie reviews (Pang and Lee, 2004). The machine translation (MT) baseline yields a sentiment classification performance on Dutch documents that is significantly inferior to the reported performance on English documents.

	Positive			Negative			Overall	
Method	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy	F_1
MT	0.416	0.385	0.400	0.428	0.460	0.443	0.423	0.422
SMap	0.547	0.500	0.523	0.540	0.587	0.562	0.543	0.542
SProp1	0.428	0.397	0.412	0.438	0.470	0.453	0.433	0.433
SProp2	0.596	**0.582**	0.589	0.591	0.605	0.598	0.593	0.593
SProp3	**0.633**	0.578	**0.605**	**0.612**	**0.665**	**0.637**	**0.622**	**0.621**

Table 4.2: The 10-fold cross-validated performance of our considered methods of classifying the sentiment conveyed by our collection of opinionated Dutch documents by exploiting an existing sentiment analysis approach for English documents. The best performance is printed in bold for each performance measure.

Benchmark	MT	SMap	SProp1	SProp2	SProp3
MT	0.000	0.286***	0.026	0.404***	0.471***
SMap	-0.222***	0.000	-0.202***	0.092**	0.144**
SProp1	-0.025	0.254***	0.000	0.369***	0.435***
SProp2	-0.288***	-0.084**	-0.270***	0.000	0.048*
SProp3	-0.320***	-0.126**	-0.303***	-0.046*	0.000

Table 4.3: Relative differences of the 10-fold cross-validated overall accuracy of our considered approaches, benchmarked against one another on our collection of opinionated Dutch documents. Performance differences marked with * are statistically significant at $p < 0.05$, those marked with ** are significant at $p < 0.01$, and those marked with *** are significant at $p < 0.001$.

Benchmark	MT	SMap	SProp1	SProp2	SProp3
MT	0.000	0.286***	0.026	0.407***	0.473***
SMap	-0.223***	0.000	-0.203***	0.094**	0.145**
SProp1	-0.025	0.254***	0.000	0.372***	0.436***
SProp2	-0.289***	-0.086**	-0.271***	0.000	0.047
SProp3	-0.321***	-0.126**	-0.303***	-0.045	0.000

Table 4.4: Relative differences of the 10-fold cross-validated macro-level F_1-score of our considered approaches, benchmarked against one another on our collection of opinionated Dutch documents. Performance differences marked with * are statistically significant at $p < 0.05$, those marked with ** are significant at $p < 0.01$, and those marked with *** are significant at $p < 0.001$.

When using the MT method, we obtain an overall accuracy and macro-level F_1-score of about 47%. The SMap method yields an overall sentiment classification accuracy and macro-level F_1-score of about 54% on Dutch documents, whereas these scores amount to about 62% for the SProp method.

The experimental results on our corpus of Dutch documents show that our novel cross-lingual sentiment score mapping method (SMap) significantly outperforms our machine translation (MT) baseline with about 29% (in relative terms), caused by increased precision and recall for both positive and negative documents. Clearly, valuable information on sentiment is (partially) contained by the semantics of our source language (i.e., English), and is as such preserved when accounting for these semantics by mapping the sentiment lexicon to our target language (i.e., Dutch) through relations between language-specific semantic lexicons. Accounting for semantics when propagating the sentiment of a seed set of sentiment-carrying words within a language (SProp) has even greater potential than exploiting semantics when mapping sentiment across languages. SProp significantly outperforms both MT and SMap with up to about 47% and 14%, respectively, in relative terms. This suggests that sentiment is not only linked to word meanings, but tends to be language-specific as well.

The MT approach may be thwarted by text meaning getting lost in translation. With the SMap method, noise may be introduced on word-level meanings, which apparently do not only depend on semantics, but can be language-specific as well. The SProp method is insensitive to such translation errors, as it depends on language-specific seed sets of sentiment-carrying words. The advantage of SProp does however appear to depend on the set of seed words used in the sentiment lexicon creation process. Our results suggest a sensitivity of the sentiment classification performance to the size of the seed set.

The smallest seed set, i.e., the seed set used by SProp1, does not yield significant improvements over any of our considered methods. Conversely, a somewhat larger seed set, i.e., the set used by SProp2, yields significant improvements over the MT baseline and the SProp1 method, as well as a small, yet significant improvement over the SMap approach. The seed set used by SProp3, i.e., the largest seed set, yields the largest, mostly significant improvements over the MT, SMap, SProp1, and SProp2 approaches. This may be explained by a larger part of the sentiment lexicon being manually annotated (i.e., the sentiment-carrying words in the seed sets), as well as by such larger initial lexicons being expanded to even larger sentiment lexicons.

The comparably large SProp 2 and SProp 3 sentiment lexicons are clearly visible in Figure 4.4, which schematically represents the coverage of the terms in DutchWordNet (Cornetto) by those in our Dutch corpus and our generated Dutch sentiment lexicons.

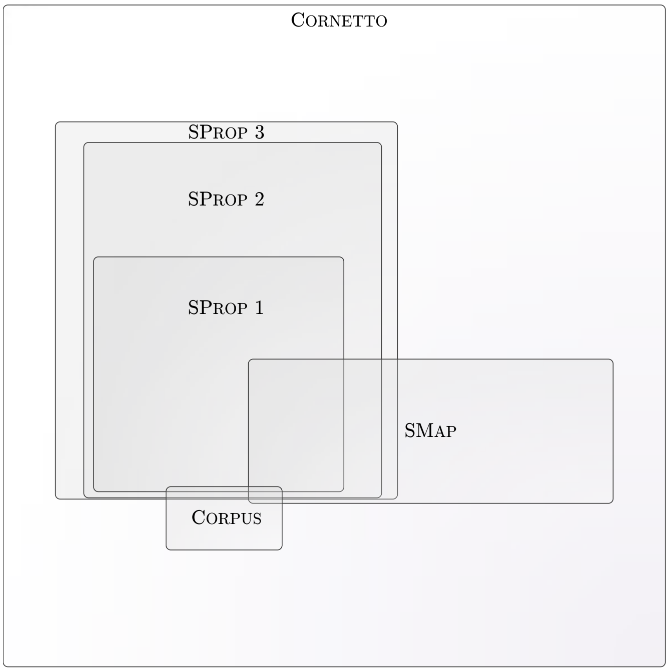


Figure 4.4: Coverage of the terms (i.e., unique combinations of lemmas with their parts-of-speech) in DutchWordNet (Cornetto) by those occurring in our Dutch documents, and by the *sentiment-carrying* terms in the Dutch sentiment lexicons generated by our SMap and SProp methods.

Figure 4.4 shows that SProp 1, SProp 2, and SProp 3 respectively cover 13%, 24%, and 29% of all terms in DutchWordNet (Cornetto), while covering 8%, 18%, and 20%, respectively, of all terms in our corpus. Interestingly, the SMap lexicon yields a significantly better performance than the SProp 1 lexicon, even though SMap only covers about 8% of the words in the corpus as well (albeit a different subset). Moreover, while having double the coverage of corpus terms of SMap, SProp 2 significantly outperforms SMap with only about 9%. Hence, the sentiment-carrying words in the SMap lexicon, constructed by exploiting semantic relations between languages, are comparably valuable in the analysis of the sentiment conveyed by our Dutch documents. This suggests that not only the size, but also the suitability of the seed sets for the corpus matters.

Figure 4.4 additionally shows that the SProp lexicons mostly cover a different part of the terms in DutchWordNet (Cornetto) than the SMap lexicon does. Especially the larger SProp lexicons cover a large part of the space, in addition to the 24%, 35%, and 40% coverage of the SMap lexicon by the respective SProp 1, SProp 2, and SProp 3 lexicons. The extra coverage of the larger SProp lexicons helps improve their performance over the SMap lexicon. This confirms the importance of exploiting semantic relations within a language when constructing a sentiment lexicon.

A failure analysis has revealed that the SProp approach occasionally fails where SMap succeeds. This tends to happen when analyzing the sentiment conveyed by texts containing sentiment-carrying words that have not been assigned appropriate scores in the sentiment score propagation process. An appropriate sentiment score may not have been assigned because either the associated synset was not reached by the propagation process, or the sentiment score was significantly diminished because of the synset being too far away from the (possibly non-optimal) seed words. Additionally, we have encountered cases in which the SMap method fails, where the SProp variants succeed. This happens when the SMap mappings do not capture the true semantics of words in Dutch, whereas the propagated SProp lexicons better approximate this.

Our failure analysis has additionally revealed that, occasionally, all of our methods fail because of misinterpreting texts. Such misinterpretations are typically caused by our approaches not accounting for, e.g., negation or amplification of sentiment. Additionally, sarcasm and proverbs are interpreted literally by our current methods, as they are not covered by our employed resources. Hashtags and other (misspelled) terms that are neither in our employed semantic lexicon nor in the constructed sentiment lexicons are another source of misinterpretations. Last, more complex structures of sentences, paragraphs, and documents are not currently taken into account. As these structures constitute the way in which sentiment-carrying words convey an author's sentiment, not accounting for these structures can cause a misinterpretation of the text in terms of its conveyed sentiment.

4.5 Conclusions

In this chapter, we have explored several methods of expanding an existing lexicon-based sentiment analysis approach for a reference language, i.e., English, to another language, i.e., Dutch. Our experimental results suggest that, when analyzing the sentiment conveyed by texts in the target language, we cannot rely on an existing, well-performing sentiment lexicon for the reference language when simply machine-translating texts to the reference language and subsequently using the existing sentiment analysis method in order to classify the sentiment of the translated texts. Conversely, when we map sentiment from the well-performing sentiment lexicon for our reference language to the target language by exploiting relations between language-specific semantic lexicons, we can achieve significantly better sentiment classification performance in the target language. Accounting for semantics by propagating sentiment of a seed set of sentiment-carrying words to semantically related words within the target language has even greater potential, provided that the seed set of sentiment-carrying words is sufficiently large.

As such, our findings indicate that sentiment is linked to semantics and, moreover, tends to have a language-specific dimension. This suggests that semantics could be exploited within a language, in addition to their use as universal link between languages when constructing sentiment lexicons in a target language. Nevertheless, our novel sentiment mapping method, exploiting relations between language-specific semantic lexicons, has two attractive advantages over our considered language-specific sentiment propagation method. First, in order for sentiment propagation to be truly effective, a large set of seed words in the target language is needed, whereas our sentiment mapping method does not need a seed set at all. Second, the sentiment propagation method is computationally more complex than the sentiment mapping method.

All in all, one of the key insights brought forward by our work is that, in order to be able to exploit the full potential of sentiment analysis in real-life decision support systems by supporting natural language content in multiple languages, semantic relations between and within languages should be carefully considered. With the accuracy levels that can be obtained by our semantics-guided methods, sentiment-related information that is extracted from text in other languages than the reference language can be presented to decision makers as a rough indication of where their attention may be needed.

Our findings warrant several directions for future work. First, we could validate our findings for another target language. Another possible direction for future research could be to further optimize the seed sets used for the sentiment propagation process, such that they, e.g., maximize the coverage of the resulting semantic lexicon. Another direction for future work could be to explore how to combine the sentiment propagation process with our proposed semantics-guided cross-lingual sentiment mapping approach in order to best exploit the strengths of both approaches. Last, as our findings indicate that sentiment is only partly conveyed by semantics, an important direction for future research would be to explore other cues, e.g., the (rhetorical) roles that words play in a text, in order to better understand the sentiment that people intend to convey. This line of research is further explored in the following chapters of this dissertation.

Chapter 5

Using the Rhetorical Structure of Text in Sentiment Analysis[*]

One of the key open research issues in automated sentiment analysis lies in dealing with structural aspects of text when analyzing its conveyed sentiment. Recent work uses structural aspects of text in order to distinguish important text segments from less important ones in terms of their contribution to the overall sentiment. However, existing methods are confined to using shallow analyses of textual structure for making coarse-grained distinctions between text segments. These methods do not account for a text's fine-grained, hierarchical rhetorical structure. We hypothesize that a better understanding of a text's associated sentiment can be obtained by guiding automated sentiment analysis by the full rhetorical structure of text. We evaluate our hypothesis in a lexicon-based sentiment analysis framework that applies the Rhetorical Structure Theory at the level of sentences, paragraphs, and documents. On an English movie review corpus, we obtain significant polarity classification performance improvements compared to baselines not or only shallowly accounting for rhetorical structure, with the best results generated by exploiting a text's full sentential rhetorical structure.

[*] This chapter is based on the articles "B. Heerschop, F. Goossen, A. Hogenboom, F. Frasincar, U. Kaymak, and F. de Jong. Polarity Analysis of Texts using Discourse Structure. In *20th ACM Conference on Information and Knowledge Management (CIKM 2011)*, pages 1061–1070. Association for Computing Machinery, 2011." and "A. Hogenboom, F. Frasincar, F. de Jong, and U. Kaymak. Using Rhetorical Structure in Sentiment Analysis. *Communications of the ACM*, 58(7):69–77, 2015."

5.1 Introduction

The Web allows users to express opinions on just about anything through an ever-increasing amount of short messages, blog posts, or reviews. Automated sentiment analysis techniques can extract traces of people's sentiment – i.e., people's attitude towards certain topics – from such texts (Feldman, 2013). This can yield competitive advantages for businesses (Arnold and Vrugt, 2008; Bal et al., 2011; Ludvigson, 2004; O'Hare et al., 2009). Thus, identifying people's opinions on such topics (Kim and Hovy, 2004), or identifying the pros and cons of products (Kim and Hovy, 2006) are promising applications for sentiment analysis tools. Sentiment analysis can be of added value in other domains too, e.g., in the political domain (Mullen and Malouf, 2006).

A well-studied problem in the field of automated sentiment analysis is the classification of the polarity of natural language text. Typical lexicon-based methods rely on the occurrence of sentiment-carrying words, listed in a sentiment lexicon. As many commercial systems use such essentially simplistic sentiment analysis techniques, Feldman (2013) pleads for finding satisfactory solutions to several open research issues. One of the key issues lies in dealing with a text's structure when analyzing its conveyed sentiment. Structural aspects may contain valuable information (Devitt and Ahmad, 2007; Hogenboom et al., 2010b; Lioma et al., 2012; Marcu, 2000; Pang et al., 2002; Taboada et al., 2008). For instance, sentiment-carrying words in a conclusion may contribute more to the overall sentiment of a text than sentiment-carrying words in, e.g., background information do.

Recent lexicon-based polarity classification methods use structural aspects of text in order to distinguish important text segments from less important ones in terms of their contribution to a text's overall sentiment. The sentiment conveyed by text segments is typically weighted in accordance with the importance of these segments when determining a text's overall sentiment. In early work, a segment's importance was related to its position in a text (Mao and Lebanon, 2006; Pang et al., 2002). More recent work makes coarse-grained distinctions between text segments based on their lexical cohesion (Devitt and Ahmad, 2007) or rhetorical roles (Taboada et al., 2008).

Commonly, such rhetorical roles are identified by applying the Rhetorical Structure Theory (RST) as proposed by Mann and Thompson (1988). In existing RST-based sentiment analysis approaches (Taboada et al., 2008), text is typically segmented in accordance with the top-level splits of (mostly sentence-level) RST trees. Even though Taboada et al. (2008) have already demonstrated the potential of exploiting such isolated rhetorical *relations* in sentiment analysis, the full rhetorical *structure* in which these relations are defined has up to this point been ignored.

However, any text segment that is assigned a rhetorical role can consist of smaller subordinate text segments that are rhetorically related to one another, thus forming a hierarchical rhetorical tree structure. Existing RST-based polarity classification methods fail to address this issue and do not account for a text's full rhetorical structure. Such methods thus inaccurately interpret the text – an important text segment may in fact contain less important parts, or vice versa. Therefore, we hypothesize that a better understanding of a text's conveyed sentiment with respect to an entity of interest can be obtained by guiding sentiment analysis by a deep analysis of a text's rhetorical structure.

The contribution of our work is three-fold. First, as an alternative to existing, shallow RST-guided analyses that typically focus on rhetorical relations in top-level splits of RST trees, we propose to focus on the leaf nodes of RST trees or, alternatively, to account for the full RST trees. A second contribution of our work lies in our novel RST-based weighting schemes, which are more refined than existing weighting schemes (Taboada et al., 2008). Third, whereas existing work mostly guides sentiment analysis by sentence-level analyses of rhetorical structure (if at all), we additionally incorporate paragraph-level and document-level analyses of rhetorical structure into the process. We thus account for rhetorical relations across sentences and paragraphs.

The remainder of this chapter is structured as follows. In Section 5.2, we review the state-of-the-art in automated sentiment analysis, with a focus on how existing work typically exploits structural aspects of text in this process. Then, we propose and evaluate our novel approach to sentiment analysis guided by a deep analysis of the rhetorical structure of text in Sections 5.3 and 5.4, respectively. Last, we conclude and provide directions for future research in Section 5.5.

5.2 Structure-Guided Sentiment Analysis

Some existing polarity classification methods account for structural aspects of content by distinguishing text segments that are important for conveying a text's sentiment from less important ones, and by weighting the sentiment conveyed by these text segments accordingly when determining a text's overall polarity. The importance of text segments was at first assumed to be related to their absolute position in a text (Mao and Lebanon, 2006; Pang et al., 2002). However, a recent lexicon-based method bases the importance of text segments on the lexical cohesion of the text as a whole (Devitt and Ahmad, 2007) – with limited success. In a more successful approach, information on the importance of text segments is harvested from the rhetorical structure of text, as identified by applying RST (Mann and Thompson, 1988) on sentence level (Taboada et al., 2008).

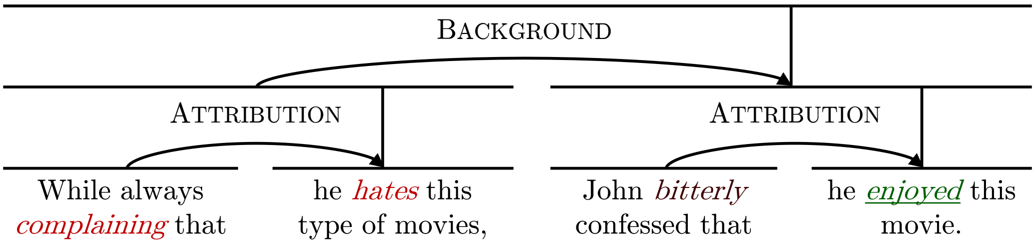

Background
Attribution
Attribution
While always *complaining* that
he *hates* this type of movies,
John *bitterly* confessed that
he *enjoyed* this movie.

Figure 5.1: A positive RST-structured sentence, consisting of nuclei (marked with vertical lines) and satellites. Negative words are printed red, in italics, whereas positive words are underlined and printed green, in italics. Sentiment-carrying words with a relatively high intensity are brighter. Horizontal lines signal the spans of the RST elements on each level of the hierarchical rhetorical structure. Arrows and their (capitalized) captions represent the relations of satellite elements to nucleus elements.

5.2.1 Rhetorical Structure Theory

RST (Mann and Thompson, 1988) is a popular discourse analysis framework. It can be used to split texts into segments that are rhetorically related to one another. Each segment may in turn be split as well, thus yielding a hierarchical rhetorical structure. Each segment is classified as either a nucleus or a satellite. Nuclei form the core of a text, whereas satellites support the nuclei and are considered less important for understanding a text. A total of 23 types of relations exist between RST elements. A satellite may, e.g., form an elaboration on or a contrast with matters presented in a nucleus.

For an example of an RST-structured sentence, let us consider the positive sentence "*While always complaining that he hates this type of movies, John bitterly confessed that he enjoyed this movie.*", which contains mostly negative words. RST can be used to describe the hierarchical rhetorical structure of its text segments, depicted in Figure 5.1. The top-level nucleus contains the core message ("*John bitterly confessed that he enjoyed this movie.*"), whereas a top-level satellite provides background information. This background satellite consists of a nucleus ("*he hates this type of movies,*") and an attributing satellite ("*While always complaining that*"). Similarly, the top-level nucleus is split into a nucleus ("*he enjoyed this movie.*") and an attributing satellite ("*John bitterly confessed that*").

When analyzing the example sentence without accounting for its structure, the relatively high frequency of negative words could advocate a negative classification. However, the sentence conveys a positive sentiment towards the movie because of the way in which the sentiment-carrying words are used in the sentence. The actual sentiment is conveyed by the nucleus "*he enjoyed this movie.*". Conversely, the other text segments simply introduce noise. In this light, the key to recognizing the intended sentiment of a text lies in accounting for its rhetorical structure in addition to its sentiment-carrying words alone.

5.2.2 Rhetorical Structure and Sentiment

Existing automated sentiment analysis approaches already exploit rhetorical relations, with some work (Taboada et al., 2008) relying more strongly on the rhetorical relations defined in RST than other methods (Polanyi and Zaenen, 2006; Somasundaran et al., 2009a,b). For instance, Polanyi and Zaenen (2006) propose to use simple discourse connectives – e.g., "*although*", "*however*", or "*but*" – in order to determine the rhetorical role of sentiment-carrying words, and to shift the polarity of affected sentiment-carrying words accordingly. Conversely, Somasundaran et al. (2009a,b) focus more on the underlying structure of a text, as they identify whether text segments do or do not reinforce one another in terms of polarity or opinion stance, and subsequently use these reinforcing and non-reinforcing relations in order to infer the overall polarity of a text.

The discourse-guided sentiment analysis methods proposed by Polanyi and Zaenen (2006) and Somasundaran et al. (2009a,b) are, however, not backed by a widely accepted framework for discourse analysis, such as RST. Yet, today's availability of RST parsers facilitates the applicability of RST to automated sentiment analysis. One of these parsers is the Sentence-level PArsing of DiscoursE (SPADE) tool (Soricut and Marcu, 2003), which creates an RST tree for each sentence in a text by means of a statistics-driven dynamic programming algorithm that relies on a text's syntactical structure and lexical features. Another discourse parser is the HIgh-Level Discourse Analyzer (HILDA) of Hernault et al. (2010), which applies machine-learning techniques that use lexical and syntactic features in order to parse the discourse structure of a text at document level. Such document-level RST trees capture rhetorical relations within as well as across sentences.

In their Sentiment Orientation CALculator (SO-CAL), Taboada et al. (2008) aim to classify the polarity of documents by only taking into account their most relevant segments. One of the approaches considered in SO-CAL is distinguishing top-level nuclei from top-level satellites in sentence-level RST trees. This approach has been proven to contribute to the polarity classification accuracy of SO-CAL.

Nevertheless, Taboada et al. (2008) do not yet make use of RST to its full extent, as their approach tends to focus on mostly isolated, coarse-grained rhetorical relations between text segments, obtained from the top-level splits of sentence-level RST trees. However, rhetorical relations are defined within a hierarchical rhetorical structure, modeled by RST trees. Nuclei may, e.g., contain less important satellites. Similarly, a satellite with background information may for instance consist of a nucleus and a contrasting satellite. We argue that such crucial nuances, as captured by a text's rhetorical *structure* rather than by its isolated rhetorical *relations*, should be accounted for in order to be able to accurately interpret the text and its conveyed sentiment.

Moreover, Taboada et al. (2008) merely differentiate between core elements of a text (nuclei) on the one hand, and any type of less important (satellite) element on the other hand. Yet, we hypothesize that the contribution of text segments to the overall sentiment of a document depends on their respective positions within the overall discourse structure and hence on their relation to other segments. For instance, a contrasting text segment may play a different role in conveying the overall sentiment than an elaboration on information in nuclei does. Therefore, we propose a more elaborate approach to utilizing RST in sentiment analysis by taking into account the distinct types of relations that can hold between nuclei and satellites.

Additionally, even though rhetorical relations may span across sentences, existing work (Taboada et al., 2008) merely focuses on sentence-level RST trees that capture rhetorical relations between parts of sentences. As such, existing work ignores rhetorical relations that may exist between larger units of analysis, such as (sets of) sentences or paragraphs. Analyzing the rhetorical structure of such larger units of analysis may provide more context for the rhetorical structure of their subordinate text segments, and may thus yield more accurate interpretations of texts as a whole. Therefore, we propose to incorporate paragraph-level and document-level analyses of rhetorical structure into the sentiment analysis process.

5.3 Classifying Polarity Using Rhetorical Structure

We propose to guide polarity classification by a deep RST-based analysis of a text's hierarchical rhetorical structure, rather than of its isolated rhetorical relations. In our analysis, we differentiate not only between nuclei and satellites, but between distinct types of nuclei and satellites as well. Moreover, we account for rhetorical relations within and across sentences by allowing for not only sentence-level, but also paragraph-level and document-level analyses of rhetorical structure.

5.3.1 Fine-Grained Analysis

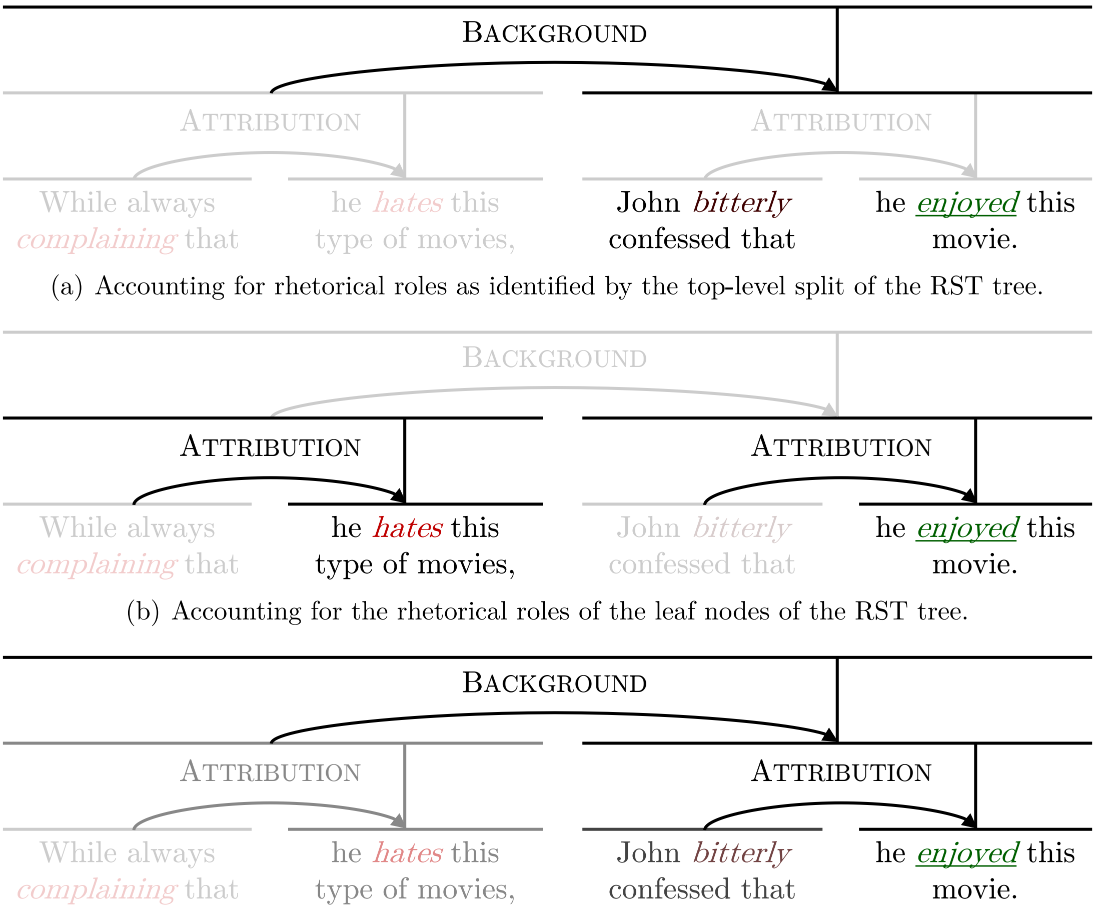
Figure 5.2 illustrates the potential of accounting for a text's rhetorical structure when classifying its polarity. When simply using the sentiment-carrying words as proxies for its overall sentiment, the example sentence introduced in Figure 5.1 in Section 5.2.1 can best be classified as a predominantly negative sentence. Accounting for the rhetorical roles of text segments as identified by the top-level split of the RST tree enables a more elaborate, but still coarse-grained analysis of the overall sentiment, as depicted in Figure 5.2(a).

The top-level nucleus contains as many positive words ("*enjoyed*") as negative words ("*bitterly*") and may therefore be classified as either positive or negative. The negative words in the top-level satellite ("*complaining*" and "*hates*") trigger a negative classification of the latter segment. The information in this satellite may however not be very relevant for the sentiment conveyed by the core of the sentence and could hence be assigned a lower weight in the analysis.

However, such a coarse-grained analysis does not capture the nuances of lower-level splits of an RST tree. For instance, the top-level nucleus of our example sentence consists of two text segments, one of which is the actual core of the sentence ("*he enjoyed this movie.*"), whereas the other contains additional, possibly less relevant information ("*John bitterly confessed that*") and should therefore be assigned a lower weight in the analysis. In this light, accounting for the rhetorical roles of the leaf nodes of an RST tree rather than the top-level splits can enable a more accurate analysis of conveyed sentiment. Figure 5.2(b) illustrates the effects of accounting for the rhetorical roles of the leaf nodes of an RST tree.

Yet, a focus on leaf nodes of RST trees alone does not account for the fact that text segments' rhetorical roles are defined within the context of the rhetorical roles of the segments that embed them. For instance, the second leaf node in our example RST tree ("*he hates this type of movies*") is a nucleus and would therefore be considered as a core part of the sentence when focusing on the leaf nodes only. However, this nucleus is in fact the core of a possibly irrelevant satellite containing background information. The full rhetorical structure should be considered in the analysis in order to account for this phenomenon. Figure 5.2(c) shows that accounting for the full, hierarchical rhetorical structure of our example sentence stresses the sentiment conveyed by the fourth leaf node ("*he enjoyed this movie.*"), i.e., the nucleus of the nucleus of the sentence, while putting less emphasis on the other segments.

In order to evaluate our novel ideas, we propose a lexicon-based framework for Polarity Analysis of Text guided by its Structure (PATHOS). This framework can perform an analysis of the rhetorical structure of a text at various levels of granularity and can subsequently use this information in order to classify the text's overall polarity. Our framework differs from existing work in that it is able to perform not only shallow analyses of top-level splits of RST trees, but deeper analyses of leaf nodes and the full RST trees as well. Additionally, we differentiate between distinct types of nuclei and satellites. Last, we support the incorporation of analyses of not only sentence-level, but also paragraph-level and document-level RST trees into the sentiment analysis process. Our framework, visualized in Figure 5.3, takes several steps in order to classify the polarity of a document.


Background
Attribution
Attribution
While always *complaining* that
he *hates* this type of movies,
John *bitterly* confessed that
he *enjoyed* this movie.
Background
Attribution
Attribution
While always *complaining* that
he *hates* this type of movies,
John *bitterly* confessed that
he *enjoyed* this movie.
Background
Attribution
Attribution
While always *complaining* that
he *hates* this type of movies,
John *bitterly* confessed that
he *enjoyed* this movie.

(a) Accounting for rhetorical roles as identified by the top-level split of the RST tree.

(b) Accounting for the rhetorical roles of the leaf nodes of the RST tree.

(c) Accounting for the full, hierarchical rhetorical structure.

Figure 5.2: RST-guided interpretations of a positive sentence, consisting of nuclei (marked with vertical lines) and satellites. Negative words are printed red, in italics, whereas positive words are underlined and printed green, in italics. Sentiment-carrying words with a relatively high intensity are brighter. Horizontal lines signal the spans of the RST elements on each level of the hierarchical rhetorical structure. Arrows and their (capitalized) captions represent the relations of satellite elements to nucleus elements. Text segments and RST elements that are assigned a relatively low weight in the analysis of the conveyed sentiment are more transparent than those that receive higher weights.

5.3.2 Word-Level Sentiment Scoring

The first step in our lexicon-based binary polarity classification framework involves preprocessing a plain-text document. To this end, we first split the document into paragraphs and, subsequently, sentences and words. Then, for each sentence, the Part-of-Speech (POS) and lemma of each word is determined. Based on the identified POS and lemma, the word sense of each word is then disambiguated by using a Lesk-based algorithm (Lesk, 1986) that selects the sense with the highest semantic similarity to the word's context, as measured by means of a similarity function proposed by Baazaoui Zghal et al. (2007).

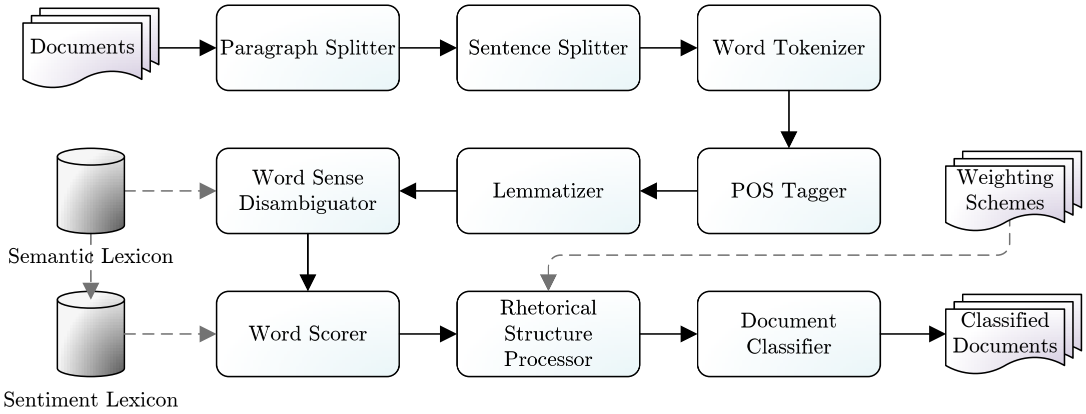


Figure 5.3: A schematic overview of PATHOS, our lexicon-based binary polarity classification framework. Solid arrows signal the information flow, whereas dashed arrows indicate a used-by relationship.

Our applied word sense disambiguation algorithm is an unsupervised algorithm that can adequately determine senses in a relatively small amount of time. This renders it an attractive alternative to other unsupervised algorithms like the SSI (Navigli and Velardi, 2005) and the original Lesk (Lesk, 1986) algorithms, which are typically slower. In our algorithm (described in Algorithm 5.1), we first retrieve all possible senses for a word t, given its POS. Then, the similarity $\varphi\left(Z_\omega, \Gamma\right)$ of a set Z_ω – i.e., the semantic neighborhood of a word sense ω – with the ambiguous word's context Γ – denoting its lexical neighborhood within a sentence – is defined as the number of overlapping terms, i.e.,

$$\varphi\left(Z_\omega, \Gamma\right) = \left|Z_\omega \cap \Gamma\right|, \tag{5.1}$$

where Z_ω is a set containing the words in the WordNet (Fellbaum, 1998) synset for word sense ω, the words in this synset's gloss (i.e., description), and all of the synset's synonyms, hyponyms, and hypernyms. Furthermore, Γ contains all the words in the sentence, except for the word t. The set Z_{ω^*} which has the highest similarity to Γ is selected, and thus gives the most similar sense ω^*. Following Baazaoui Zghal et al. (2007), we select the set which has the highest number of element, if there are more sets with the same similarity.

After having determined each word's POS, lemma, and word sense, the sentiment associated with these combinations of POS, lemma, and word sense can be retrieved from a sentiment lexicon, e.g., SentiWordNet 3.0 (Baccianella et al., 2010). The word-level sentiment scores thus obtained are subsequently weighted in accordance with the identified rhetorical role of their associated text segments. The resulting scores are then aggregated in order to determine the sentiment score of a document, which in turn is used in order to classify its polarity. These steps are detailed in Sections 5.3.3, 5.3.4, and 5.3.5.

Algorithm 5.1: Word sense disambiguation.

```
input  : The word $t$ to be disambiguated, its POS $pos$, and its context $\Gamma$
output : The sense $\omega^*$ of $t$ with the highest semantic similarity to the words in the
         context $\Gamma$
$\Omega$ = getSenseSynsets($t$, $pos$);
$\omega^* = \emptyset$;
$Z_{\omega^*} = \emptyset$;
$\varphi(Z_{\omega^*}, \Gamma) = -\infty$;
if $|\Omega| \leq 1$ then
    return $\Omega$;
else
    foreach $\omega \in \Omega$ do
        $Z_\omega$ = {getWords($\omega$), getGlossWords($\omega$)};
        $Z_\omega$ = {$Z_\omega$, getSynonyms($\omega$), getHyponyms($\omega$), getHypernyms($\omega$)};
        $\varphi(Z_\omega, \Gamma) = |Z_\omega \cap \Gamma|$;
        if $(\varphi(Z_\omega, \Gamma) > \varphi(Z_{\omega^*}, \Gamma)) \mid\mid (((\varphi(Z_\omega, \Gamma) == \varphi(Z_{\omega^*}, \Gamma))$ && $(|Z_\omega| > |Z_{\omega^*}|))$ then
            $\omega^* = \omega$;
            $Z_{\omega^*} = Z_\omega$;
            $\varphi(Z_{\omega^*}, \Gamma) = \varphi(Z_\omega, \Gamma)$;
        end
    end
    return $\omega^*$;
end
```

5.3.3 Rhetorical Structure Processing

In order to account for the rhetorical structure of a document d when determining its polarity, first of all, sentiment scores are computed for each identified text segment s_i. Our framework supports several ways of computing these segment-level sentiment scores, which depend on the way in which the rhetorical structure of a document is taken into account in the sentiment analysis process. These supported methods include a baseline method (formalizing Figure 5.1), as well as several RST-based methods (formalizing Figure 5.2) that can be applied to sentence-level, paragraph-level, and document-level RST trees.

5.3.3.1 Baseline

As a baseline, we consider text segments to be the sentences S_d of document d, with their associated baseline sentiment score $\zeta_{s_i}^B$ being the weighted sum of each word t_j's sentiment score ζ_{t_j} and its associated weight w_{t_j} that quantifies its importance, i.e.,

$$\zeta_{s_i}^B = \sum_{t_j \in s_i} \left(\zeta_{t_j} \cdot w_{t_j} \right), \quad \forall s_i \in S_d. \tag{5.2}$$

5.3.3.2 Top-Level Rhetorical Structure

As an alternative to the baseline method, our framework supports the top-level RST-based sentiment analysis approach applied in existing work discussed in Section 5.2.2, on sentence, paragraph, and document level. We refer to this approach as T. Here, the sentiment score $\zeta_{s_i}^T$ of a top-level RST segment s_i is defined as the sum of the sentiment ζ_{t_j} associated with each word t_j in segment s_i, weighted with a weight w_{s_i} associated with the rhetorical role of the segment, i.e.,

$$\zeta_{s_i}^T = \sum_{t_j \in s_i} \left(\zeta_{t_j} \cdot w_{s_i} \right), \quad \forall s_i \in T_d, \tag{5.3}$$

with T_d representing all top-level RST nodes in the RST trees for document d.

5.3.3.3 Leaf-Level Rhetorical Structure

Another method supported by our framework is our novel leaf-level RST-based approach L. The sentiment score $\zeta_{s_i}^L$ of an RST segment s_i from the leaf nodes L_d of a sentence-level, paragraph-level, or document-level RST tree for document d is computed as the sum of the sentiment score of its words, weighted for the segment's rhetorical role, i.e.,

$$\zeta_{s_i}^L = \sum_{t_j \in s_i} \left(\zeta_{t_j} \cdot w_{s_i} \right), \quad \forall s_i \in L_d. \tag{5.4}$$

5.3.3.4 Hierarchical Rhetorical Structure

The last supported approach is our novel method of accounting for the full path from an RST tree root to a leaf node, such that the sentiment conveyed by the latter is weighted while accounting for its rhetorical context. In our hierarchy-based sentiment scoring method H, we model the sentiment score $\zeta_{s_i}^H$ of a leaf-level RST segment s_i as a function of the sentiment scores of its words and the weights w_{r_n} associated with the rhetorical role of each node r_n from the nodes P_{s_i} on the path from the root to the leaf, i.e.,

$$\zeta_{s_i}^H = \sum_{t_j \in s_i} \zeta_{t_j} \cdot \left(\frac{\sum_{r_n \in P_{s_i}} \left(|w_{r_n}| \cdot \delta^{-(\lambda_{r_n} - 1)} \right)}{\sum_{r_n \in P_{s_i}} \delta^{-(\lambda_{r_n} - 1)}} \right) \cdot \prod_{r_n \in P_{s_i}} \operatorname{sgn}(w_{r_n}), \quad \forall s_i \in L_d, \quad \delta > 1, \tag{5.5}$$

where δ represents a diminishing factor and λ_{r_n} signals the level of node r_n in the RST tree, where the level of the root node equals 1. The sentiment score $\zeta_{s_i}^H$ of a leaf-level RST segment s_i is essentially the product of the sentiment scores of its words and an RST-based weight of the segment. The latter weight is the product of two components.

The first component is the weighted average of the intensity (i.e., absolute values) of the weights w_{r_n} associated with the rhetorical role of each node r_n from the nodes P_{s_i} on the path from the root to the leaf node, where the contribution of each level in the path depends on a diminishing factor δ. For $\delta > 1$, each subsequent level contributes less than its parent does to the weight, thus preserving the hierarchy of the relations in the path. The normalization of the intensity ensures comparability of $\zeta_{s_i}^H$ values across segments.

The second component of a leaf-level RST segment's weight is the product of the signs of the weights of the rhetorical roles on the path from the root to the leaf node. This ensures that an odd number of negative weights in the path yields a negative weight, an even number of negative weights yields a positive weight, and a weight of 0 in the path yields a weight of 0 for the path as a whole. The assumption here is that a positive (negative) polarity turns negative (positive) when inverted, and positive (negative) when inverted again, analogous to double negation on a syntactical level in the English language.

5.3.4 Classifying Document Polarity

Having computed the sentiment for each identified text segment s_i of document d, the segment-level sentiment scores can be aggregated in order to determine the overall polarity of the document. The sentiment scores ζ_d^B, ζ_d^T, ζ_d^L, and ζ_d^H for document d are defined as

$$\zeta_d^B = \sum_{s_i \in S_d} \zeta_{s_i}^B, \tag{5.6}$$

$$\zeta_d^T = \sum_{s_i \in T_d} \zeta_{s_i}^T, \tag{5.7}$$

$$\zeta_d^L = \sum_{s_i \in L_d} \zeta_{s_i}^L, \tag{5.8}$$

$$\zeta_d^H = \sum_{s_i \in L_d} \zeta_{s_i}^H. \tag{5.9}$$

The resulting document-level sentiment scores can be used to classify document d's polarity c_d. For any of our considered structure-based sentiment scoring approaches, we classify a document as negative (-1) if its computed document-level sentiment score is negative, and as positive (1) in all other cases, i.e.,

$$c_d = \begin{cases} -1 & \text{if } (\zeta_d - \epsilon) < 0, \\ 1 & \text{otherwise}, \end{cases} \tag{5.10}$$

with ϵ being an offset correcting a possible bias in the sentiment scores caused by people's tendency to write negative reviews with rather positive words (Taboada et al., 2008).

The offset can be computed by taking the average sentiment scores of both positive and negative documents in a training set and subsequently computing the equidistant point of these scores (Taboada et al., 2008), i.e.,

$$\epsilon = 0.5 \left(\frac{\sum_{d \in \Phi} \zeta_d}{|\Phi|} + \frac{\sum_{d \in N} \zeta_d}{|N|} \right), \tag{5.11}$$

with Φ and N denoting the respective subsets of positive and negative documents in a training set.

5.3.5 Weighting Schemes

We consider six different weighting schemes for words or text segments in the polarity classification process. Two of these schemes serve as baselines and as such are applicable to the baseline sentiment scoring approach as defined in (5.2) and (5.6). The other schemes apply to any of our RST-based sentiment scoring approaches as defined in (5.3) and (5.7), in (5.4) and (5.8), and in (5.5) and (5.9).

The first scheme, i.e., the BASELINE scheme, serves as an absolute baseline and assigns each word a weight equal to 1, such that structural aspects of content are not accounted for at all. As a second baseline, we consider a position-based weighting scheme. In this POSITION scheme, word weights are uniformly distributed and range from 0 for the first word to 1 for the last word of a text, as an author's views are likely to be summarized near the end of a text (Pang et al., 2002).

Two of our considered RST-specific weighting schemes consider sentiment conveyed by nuclei to be important, and information found in satellite elements to be tangential or even irrelevant with respect to a text's overall sentiment. One scheme (I) assigns a weight of 1 to nuclei and a weight of 0 to satellites (Taboada et al., 2008). Our second scheme (II) matches the second set of weights for nuclei and satellites used by Taboada et al. (2008), i.e., 1.5 and 0.5, respectively. In both schemes I and II, we set the diminishing factor δ for the RST analysis method H to 2, such that each level in a tree is at least as important as all of its subsequent levels combined, thus enforcing a strict hierarchy.

Another considered RST-specific weighting scheme is our novel, extended weighting scheme X, in which we differentiate between nuclei and various types of satellites. By doing so, we account for the possibility of some satellite relation types contributing differently to the overall sentiment of a text than others. Additionally, we propose an extension of the X weighting scheme, in which we not only differentiate satellite weights, but also nucleus weights by their RST relation type. We refer to this full weighting scheme as F.

In both X and F, we consider both positive and negative weights, as some text segments (e.g., contrasting ones) may contribute negatively to the overall sentiment of a text. In order to allow for intensification of sentiment, we assume the weights to be in the range $[-2, 2]$. The weights and the diminishing factor δ can be optimized by means of, e.g., a genetic algorithm (Heerschop et al., 2011a) or particle swarm optimization (Chenlo et al., 2013), in order to find a solution that maximizes document polarity classification performance, assessed in terms of the macro-level F_1-score on a training set.

5.4 Evaluation

The variants of our polarity classification approach guided by the full rhetorical structure of text, discussed in Section 5.3, are evaluated by means of a set of experiments. For this purpose, we focus on a collection of 1,000 positive and 1,000 negative English movie reviews (Pang and Lee, 2004).

5.4.1 Experimental Setup

In order to be able to compare the performance of our considered polarity classification methods on our data, we have created a Java implementation of our proposed framework. We detect paragraphs using the `<P>` and `</P>` tags in the original HTML files of the reviews, as these tags signal the respective starts and ends of paragraphs. In order to segment the identified paragraphs into sentences, we rely on the preprocessing done by Pang and Lee (2004), which has resulted in sentences being separated by means of line breaks. In order to segment the identified sentences into words, we use the Stanford Tokenizer (Manning et al., 2010). For POS tagging and lemmatization, we use the OpenNLP (Baldridge and Morton, 2004) POS tagger and the JWNL API (Walenz and Didion, 2008), respectively.

Our framework relies on a semantic lexicon and a sentiment lexicon. The word senses in our framework are linked to WordNet (Fellbaum, 1998), i.e., a freely available semantic lexical resource, organized into sets of synonyms. Sentiment scores for the identified word senses are retrieved from SentiWordNet 3.0 (Baccianella et al., 2010), containing positivity, negativity, and objectivity scores for each set of synonyms available in WordNet. We use this information to compute sentiment scores for each disambiguated word by subtracting its associated negativity score from its associated positivity score, thus yielding a real number in the interval $[-1, 1]$, representing sentiment scores in the range from very negative to very positive, respectively.

Our implementation enables us to assess the performance of our considered methods of RST-guided polarity classification. We consider approaches that use sentence-level, paragraph-level, or document-level RST trees as generated by the SPADE and HILDA parsers, i.e., SPADE.S for sentence-level RST trees, and HILDA.S, HILDA.P, and HILDA.D for sentence-level, paragraph-level, and document-level RST trees, respectively. For each of these RST analysis methods, we consider the top-level RST-based polarity classifier T, our leaf-level RST-based polarity classification method L and our classifier which guides the polarity classification of a document by its full rhetorical structure, i.e., H. For all RST-based classifiers, we consider weighting schemes I, II, and our novel weighting schemes X and F. As an absolute baseline, we use the Baseline method, assuming all words to equally contribute to a text's sentiment. Additional baselines are the Position approach and two baselines representing existing work, i.e., SPADE.S T with weighting schemes I and II.

We evaluate of each of our considered methods by assessing the accuracy, precision, recall, and F_1-score on positive and negative documents, as well as by assessing the overall accuracy and macro-level F_1-score over all documents. Precision on positive (negative) documents is defined as the proportion of the positively (negatively) classified documents that have an actual classification of positive (negative). Recall of positive (negative) documents is defined as the proportion of the actual positive (negative) documents that are also classified as such. For positive and negative documents, the F_1-score is calculated as the harmonic mean of their respective precision and recall measures. The macro-level F_1-score is computed as the average of the F_1-scores of the positive and negative documents. The accuracy is defined as the overall proportion of correctly classified documents.

For our weighting schemes X and F, we optimize the weights for distinct rhetorical relations, as well as the offsets and the diminishing factor δ to be used in the RST analysis method H. We follow recent work by using a particle swarm optimization algorithm for this purpose (Chenlo et al., 2013). In this algorithm, particles search a solution space, where the coordinates of their position correspond with the weights, the offsets, and the diminishing factor δ. The fitness of a set of coordinates is modeled as the macro-level F_1-score on a training set of documents.

We apply 10-fold cross-validation in our evaluation. For each fold, we optimize the offsets, the weights for weighting schemes X and F, and the diminishing factor δ for the hierarchy-based RST-guided polarity classification method H on a set of 900 positive documents and 900 negative documents, whereas the remaining 100 positive documents and 100 negative documents are used to assess the performance of our methods on unseen data.

	Positive			Negative			Overall	
Method	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy	F_1
Baseline	0.632	0.689	0.659	0.658	0.599	0.627	0.644	0.643
Position	0.637	**0.713**	**0.673**	**0.674**	0.593	0.631	**0.653**	**0.652**
SPADE.S T I	0.638	0.675	0.656	0.655	**0.617**	0.635	0.646	0.646
SPADE.S T II	**0.640**	0.688	0.663	0.663	0.613	**0.637**	0.651	0.650

Table 5.1: Performance of our considered baseline methods, based on 10-fold cross-validation on the movie review data set. The best performance is printed in bold for each performance measure.

In our comparisons, we assess the statistical significance of observed performance differences by means of paired, one-tailed t-tests. In these tests, we compare the considered approaches against one another in terms of their mean performance measures over all ten folds. Here, the null hypothesis is that the mean performance of an arbitrary method is less than or equal to the mean performance of another method. Consequently, the alternative hypothesis is that the former method outperforms the latter.

5.4.2 Experimental Results

Our analysis consists of eight steps. First, we evaluate the performance of our baselines. Then, we evaluate the performance of our sentence-level, paragraph-level, and document-level RST-based approaches. Subsequently, we compare our considered approaches with one another, distill general patterns in performance of the methods, and identify the weakest and strongest methods for our corpus. Then, we analyze the optimized weights and diminishing factors and we demonstrate how documents are typically perceived by distinct methods. Last, we identify some caveats with respect to our findings.

5.4.2.1 Baseline Performance

The results presented in Table 5.1 demonstrate that the absolute baseline of not accounting for any structural aspects of content, i.e., Baseline, is the worst performing baseline approach on our corpus, with an overall accuracy and macro-level F_1-score of about 64%. With overall performance scores around 65%, the Position baseline method performs only slightly better than the Baseline approach. The RST-based baseline approaches exploiting top-level splits of sentence-level RST trees generated by SPADE show little improvement over the absolute baseline as well. With weighting scheme II, the performance of sentiment analysis guided by top-level splits of sentence-level RST trees is similar to the performance of the Position method.

	Positive			Negative			Overall	
Method	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy	F_1
SPADE.S T I	0.638	0.675	0.656	0.655	0.617	0.635	0.646	0.646
SPADE.S T II	0.640	0.688	0.663	0.663	0.613	0.637	0.651	0.650
SPADE.S T X	0.693	0.725	0.709	0.712	0.679	0.695	0.702	0.702
SPADE.S T F	0.703	0.726	0.715	0.717	0.694	0.705	0.710	0.710
SPADE.S L I	0.636	0.702	0.667	0.667	0.598	0.631	0.650	0.649
SPADE.S L II	0.640	0.700	0.669	0.669	0.607	0.637	0.654	0.653
SPADE.S L X	0.699	0.715	0.707	0.708	0.692	0.700	0.704	0.703
SPADE.S L F	0.705	0.731	0.718	0.721	0.694	0.707	0.713	0.712
SPADE.S H I	0.647	0.678	0.662	0.662	0.630	0.645	0.654	0.654
SPADE.S H II	0.642	0.696	0.668	0.668	0.612	0.639	0.654	0.653
SPADE.S H X	0.707	0.723	0.715	0.716	**0.700**	0.708	0.712	0.711
SPADE.S H F	**0.710**	**0.738**	**0.724**	**0.727**	0.699	**0.713**	**0.719**	**0.718**

Table 5.2: Performance of our considered sentence-level RST-based approaches utilizing the SPADE parser, based on 10-fold cross-validation on the movie review data set. The best performance is printed in bold for each performance measure.

5.4.2.2 Performance with Sentence-Level RST Trees

The performance measures of our RST-based approaches using sentence-level RST trees generated by the SPADE parser reveal interesting patterns, as can be seen in Table 5.2. First, with overall accuracy and macro-level F_1-scores around 65%, all methods utilizing weighting schemes I and II hardly outperform the baselines that do not or only naively incorporate RST-based analyses into the polarity classification process. Conversely, the extended and full weighting schemes X and F yield overall accuracy and macro-level F_1-scores of up to approximately 72%. This is considerably higher than the accuracy and macro-level F_1-scores scores of our best baselines, which are approximately 65%.

Another interesting pattern exhibited by the results in Table 5.2 is that approaches utilizing weighting scheme F typically outperform those with weighting scheme X. Additionally, our novel leaf-level RST-based polarity classification method L and especially our novel hierarchy-guided RST-based polarity classification method H typically outperform the more coarse-grained top-level RST-based approach T. Both the superiority of F over X and the superiority of H and L over T can mostly be explained by increased recall for positive documents and increased precision for negative documents.

The most successful method of sentiment analysis guided by sentential rhetorical structure as identified by means of the SPADE parser involves combining the best RST analysis method H with the best weighting scheme F, i.e., SPADE.S H F. This method yields overall accuracy and macro-level F_1-scores around 72%.

	Positive			Negative			Overall	
Method	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy	F_1
HILDA.S T I	0.633	0.676	0.654	0.652	0.608	0.629	0.642	0.642
HILDA.S T II	0.636	0.686	0.660	0.659	0.607	0.632	0.647	0.646
HILDA.S T X	0.692	0.709	0.701	0.702	0.685	0.693	0.697	0.697
HILDA.S T F	0.697	**0.745**	**0.720**	**0.726**	0.676	0.700	0.711	0.710
HILDA.S L I	0.629	0.685	0.656	0.654	0.596	0.624	0.641	0.640
HILDA.S L II	0.636	0.685	0.660	0.659	0.608	0.632	0.647	0.646
HILDA.S L X	0.698	0.711	0.705	0.706	0.693	0.699	0.702	0.702
HILDA.S L F	**0.705**	0.732	0.718	0.721	0.693	**0.707**	**0.713**	**0.712**
HILDA.S H I	0.634	0.675	0.654	0.653	0.611	0.631	0.643	0.643
HILDA.S H II	0.638	0.688	0.662	0.661	0.609	0.634	0.649	0.648
HILDA.S H X	0.699	0.693	0.696	0.695	**0.701**	0.698	0.697	0.697
HILDA.S H F	0.699	0.740	0.719	0.724	0.682	0.702	0.711	0.711

Table 5.3: Performance of our considered sentence-level RST-based approaches utilizing the HILDA parser, based on 10-fold cross-validation on the movie review data set. The best performance is printed in bold for each performance measure.

When using the HILDA parser rather than the SPADE parser for parsing the sentences in the documents in our corpus into sentence-level RST trees, the performance measures of our RST-based polarity classification methods exhibit patterns that are rather similar to those of the RST-based methods using the SPADE parser. The experimental results for our HILDA-based approaches using sentence-level RST trees are detailed in Table 5.3.

With overall accuracy and macro-level F_1-scores around 64% and 65%, weighting schemes I and II do not yield substantial improvements over our baselines that do not or only naively guide the polarity classification process by RST. On the other hand, accuracy and macro-level F_1-scores amount to about 70% for methods using weighting scheme X and to approximately 71% for those methods using weighting scheme F.

The distinction between RST analysis methods T, L, and H is less clear-cut for sentence-level RST-based methods using the HILDA parser than for those using the SPADE parser. However, small differences in terms of polarity classification performance can be observed, suggesting that the L and H methods are, to a limited extent, superior to the T approach.

The best performing HILDA-based methods that use sentence-level RST trees in order to guide the polarity classification process, are those with weighting scheme F, i.e., HILDA.S T F, HILDA.S L F, and HILDA.S H F. These methods yield accuracy and macro-level F_1-scores up to about 71%.

	Positive			Negative			Overall	
Method	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy	F_1
HILDA.P T I	0.618	0.638	0.628	0.626	0.605	0.615	0.622	0.621
HILDA.P T II	0.628	0.674	0.650	0.648	0.600	0.623	0.637	0.637
HILDA.P T X	0.681	0.697	0.689	0.690	0.674	0.682	0.686	0.685
HILDA.P T F	0.703	0.702	0.702	0.702	0.703	0.703	0.703	0.702
HILDA.P L I	0.632	0.684	0.657	0.656	0.602	0.628	0.643	0.642
HILDA.P L II	0.633	0.685	0.658	0.657	0.603	0.629	0.644	0.643
HILDA.P L X	0.690	0.705	0.697	0.698	0.683	0.691	0.694	0.694
HILDA.P L F	0.701	**0.720**	**0.710**	**0.712**	0.693	0.702	0.707	0.706
HILDA.P H I	0.583	0.609	0.596	0.591	0.565	0.578	0.587	0.587
HILDA.P H II	0.629	0.682	0.655	0.653	0.598	0.624	0.640	0.639
HILDA.P H X	0.706	0.683	0.694	0.693	0.716	0.704	0.700	0.699
HILDA.P H F	**0.713**	0.692	0.703	0.701	**0.722**	**0.711**	**0.707**	**0.707**

Table 5.4: Performance of our considered paragraph-level RST-based approaches utilizing the HILDA parser, based on 10-fold cross-validation on the movie review data set. The best performance is printed in bold for each performance measure.

5.4.2.3 Performance with Paragraph-Level RST Trees

When we guide our RST-based methods for polarity classification by paragraph-level RST trees produced by the HILDA parser, several observations can be made from the experimental results detailed in Table 5.4. The first observation that can be made is that the considered methods applying weighting scheme I do not seem to be able to outperform our baselines. The overall accuracy and macro-level F_1-scores vary from about 59% to 64% for our paragraph-level RST-based methods with weighting scheme I. Additionally, with accuracy and macro-level F_1-scores around 64%, weighting scheme II does not yield a clear advantage over our baselines either in terms of polarity classification performance.

More promising paragraph-level RST-based polarity classification approaches utilize the extended weighting scheme X or the full weighting scheme F. Weighting scheme X yields accuracy and F_1-scores of about 69% to 70%, whereas weighting scheme F yields overall performance scores of 70% to 71%. These overall performance scores are well above the scores for our baselines.

Out of the three considered RST analysis methods, i.e., T, L, and H, the comparably shallow and coarse-grained T method typically yields the worst performance on our corpus when accounting for paragraph-level RST trees in the polarity classification process. The deeper, more fine-grained analysis of the L and H methods typically yields better performance, yet the difference between L and H tends to be fairly small. The difference between the best performing methods HILDA.P H F and HILDA.P L F is negligible.

	Positive			Negative			Overall	
Method	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy	F_1
HILDA.D T I	0.627	0.616	0.621	0.622	0.633	0.628	0.625	0.624
HILDA.D T II	0.627	0.650	0.639	0.637	0.614	0.625	0.632	0.632
HILDA.D T X	0.682	0.689	0.685	0.686	0.678	0.682	0.684	0.683
HILDA.D T F	0.684	0.696	0.690	0.691	0.679	0.685	0.688	0.687
HILDA.D L I	0.627	0.679	0.652	0.650	0.596	0.622	0.638	0.637
HILDA.D L II	0.631	0.687	0.658	0.656	0.598	0.626	0.643	0.642
HILDA.D L X	0.689	0.719	0.704	0.706	0.675	0.690	0.697	0.697
HILDA.D L F	0.701	**0.727**	**0.714**	**0.717**	0.690	0.703	**0.709**	**0.708**
HILDA.D H I	0.580	0.516	0.546	0.564	0.627	0.594	0.572	0.570
HILDA.D H II	0.630	0.663	0.646	0.645	0.611	0.627	0.637	0.637
HILDA.D H X	0.706	0.696	0.701	0.700	**0.710**	0.705	0.703	0.703
HILDA.D H F	**0.707**	0.708	0.707	0.707	0.706	**0.707**	0.707	0.707

Table 5.5: Performance of our considered document-level RST-based approaches utilizing the HILDA parser, based on 10-fold cross-validation on the movie review data set. The best performance is printed in bold for each performance measure.

HILDA.P H I forms an exception to the observation of the H method outperforming the L and T methods, as HILDA.P H I yields by far the worst performance of our paragraph-level RST-based methods. The cause of this lies in the satellite relations being assigned weights of 0 in weighting scheme I. In the H method, this yields an RST-based weight of 0 for all leaf nodes with one or more satellite relations in the path from the root nodes of the associated RST trees to the respective leaf nodes. Only a small part of a text – consisting of leafs with only nucleus relations in their associated path – is thus assumed to convey relevant sentiment. This may result in a rather strong and narrow focus on specific parts of a text, whereas other parts may contain valuable information as well.

5.4.2.4 Performance with Document-Level RST Trees

The performance measures in Table 5.5 exhibit several patterns in the performance of RST-based polarity classification guided by document-level RST trees as generated by the HILDA parser. First, the most distinctive characteristic of our considered approaches appears to be the employed weighting scheme. The methods that use weighting scheme I obtain overall accuracy and macro-level F_1-scores between about 57% and 64%, whereas those using weighting scheme II perform only slightly better with performance scores of 63% to 64%. On the other hand, both weighting schemes X and F exhibit better performance, with overall performance scores ranging from about 69% to 70% for scheme X, or even 71% for scheme F. These scores convincingly exceed the scores for our baselines.

For all considered weighting schemes, the hierarchy-based RST analysis method H typically outperforms the leaf-level approach L, which in turn outperforms the top-level method T. Yet, for weighting scheme F, the H and L methods yield rather similar overall performance scores, thus making both HILDA.D L F and HILDA.D H F the best performing document-level RST-based polarity classification methods. Weighting scheme I in the HILDA.D H I method forms another exception, as its narrow focus caused by ignoring all segments with one or more satellite relations in their associated paths in the document-level RST trees yields the worst performance of all of our considered document-level RST-based polarity classification approaches.

5.4.2.5 Comparison

In order to be able to identify the most successful strategies for dealing with a document's (rhetorical) structure when classifying its polarity, we compare all 50 considered methods with one another in terms of their mean performance over ten folds. Here, the null hypothesis is that the mean performance of a method is less than or equal to the mean performance of another method, i.e., that the former method is outperformed by the latter. For all combinations of methods, Figures 5.4 and 5.5 visualize the p-values for the overall accuracy and macro-level F_1-scores, respectively.

In Figures 5.4 and 5.5, we have ordered our considered polarity classification methods from top to bottom and from left to right in accordance with an initial ranking from the worst to the best performing methods, based on the experimental results presented in Sections 5.4.2.1, 5.4.2.2, 5.4.2.3, and 5.4.2.4. To this end, we have first ordered our methods based on their weighting scheme (Baseline, Position, I, II, X, and F), then based on their level of analysis (HILDA.D, HILDA.P, HILDA.S, and SPADE.S), and finally based on their applied RST analysis method (T, L, and H).

Darker colors in Figures 5.4 and 5.5 represent lower p-values for the null hypothesis of the overall performance scores of the methods in the columns being smaller than or equal to the scores of the methods in the rows. Consequently, darker horizontal lines emerging in the plots signal weak approaches, generally outperformed by the other methods. Additionally, darker vertical lines emerging in the plots signal competitive approaches, generally outperforming the other considered approaches. Last, provided that the applied ordering of the methods is correct, one would expect darker colors to appear towards the upper right corner of both Figure 5.4 and Figure 5.5.

Three general trends can be observed in Figures 5.4 and 5.5, which both exhibit very similar patterns. First, the dark upper right quadrants indicate that weighting schemes X and F significantly outperform weighting schemes Baseline, Position, I, and II.

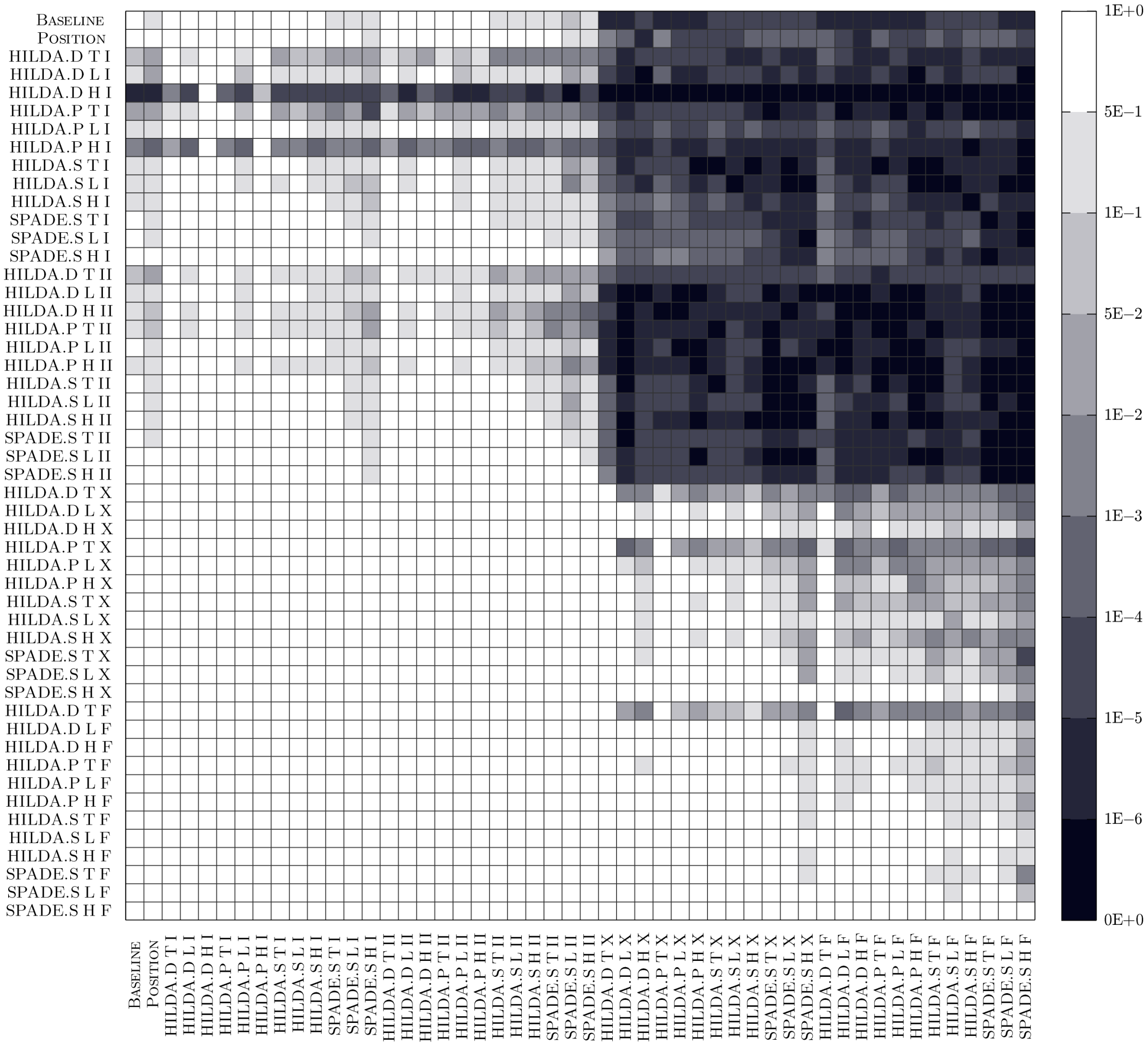


Figure 5.4: The p-values for the paired, one-tailed t-test assessing the null hypothesis of the mean accuracy of the methods in the columns being smaller than or equal to the mean accuracy of the methods in the rows.

Additionally, weighting scheme F significantly outperforms weighting scheme X in most cases, as signaled by the relatively dark top-right quadrant of the bottom-right quadrant of both Figure 5.4 and Figure 5.5. Conversely, weighting schemes Baseline, Position, I, and II do not exhibit many clearly significant differences in performance with respect to one another, as can be derived from the rather light top-right quadrants of the top-left quadrants of Figures 5.4 and 5.5.

A second trend that can be observed in Figures 5.4 and 5.5 is that RST-based polarity classification methods guided by document-level RST trees are typically outperformed by comparable methods that guide polarity classification by paragraph-level RST trees.

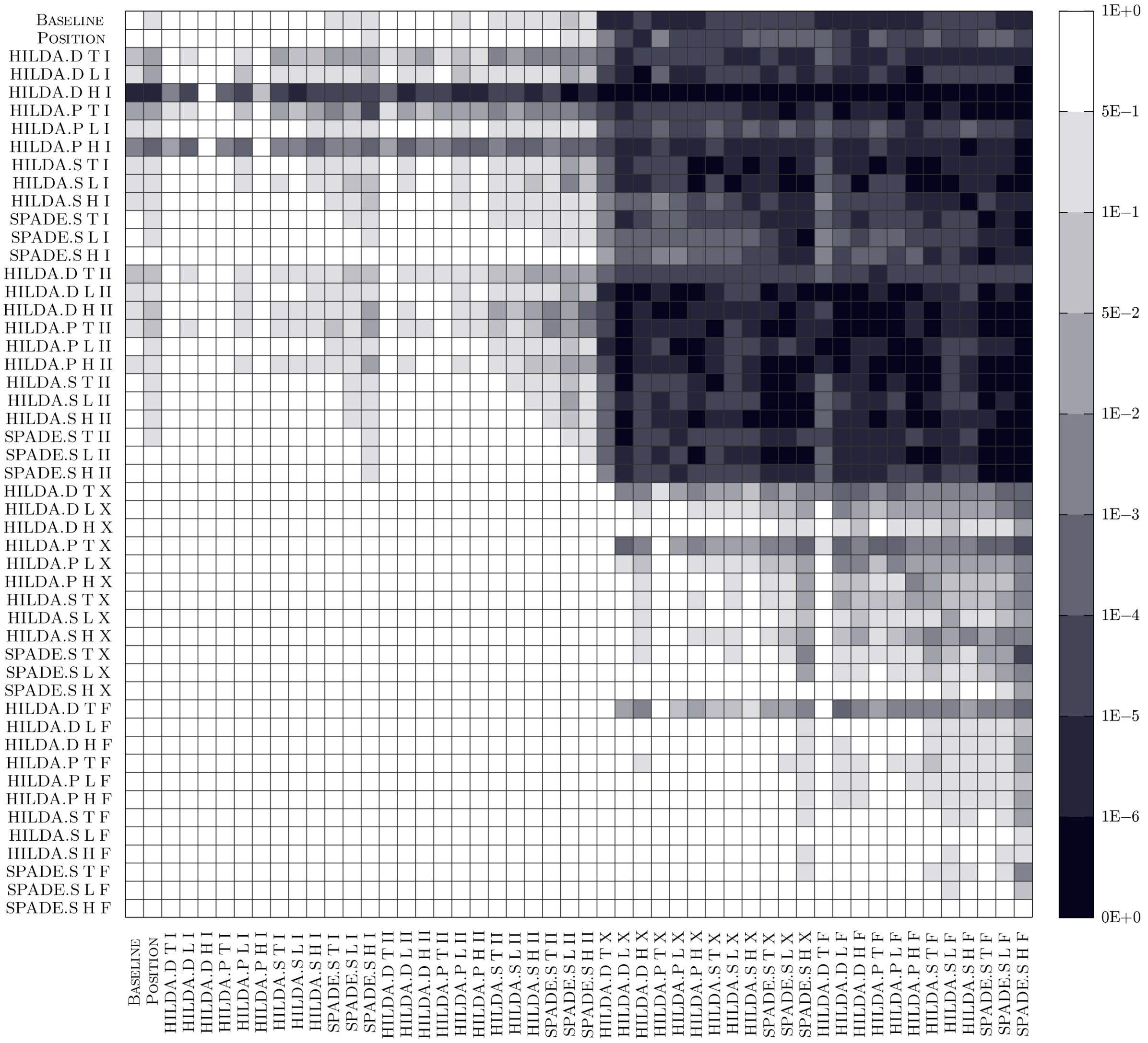


Figure 5.5: The p-values for the paired, one-tailed t-test assessing the null hypothesis of the mean macro-level F_1-scores of the methods in the columns being smaller than or equal to the mean macro-level F_1-scores of the methods in the rows.

Moreover, sentence-level RST trees appear to yield the best results. This result is somewhat counterintuitive, as one would expect a text to be best interpreted when accounting for its complete structure, rather than for the isolated structure of its smaller, subordinate units like paragraphs or sentences.

This counterintuitive result can be explained by misclassifications of rhetorical relations being potentially more harmful in larger RST trees than they are in smaller RST trees. A misclassified relation in one of the top levels of a document-level RST tree can cause a misinterpretation of a large part of the document, whereas the consequences of similar misclassifications in sentence-level RST trees are limited to single sentences only.

An additional explanation lies in the fact that especially the T analysis method is a rather coarse-grained method for analyzing larger RST trees. For document-level trees, the T method effectively differentiates between two parts of a complete document, rather than between two parts per paragraph or sentence when applying the T method to paragraph-level or sentence-level RST trees, respectively.

A third trend emerging from the visualizations in Figures 5.4 and 5.5 is that approaches applying the hierarchy-based RST analysis method H typically slightly outperform comparable approaches that use the leaf-based analysis L instead, which in turn tend to significantly outperform comparable approaches that use the top-level RST analysis method T. Clearly, the deeper analyses L and H yield a significant advantage in terms of polarity classification performance over the rather shallow analysis method T.

Apart from these three general trends, some of our 50 considered approaches stand out in particular. First, the HILDA.D T and HILDA.P T variants are relatively weak in terms of polarity classification performance, especially when using weighting schemes I and II. The top-level RST analysis method T results in a document being segmented in only two parts when using document-level RST trees, and two parts per paragraph when using paragraph-level RST trees. The combination of such relatively coarse-grained segmentations with rather naive weighting schemes is the root of the comparably weak performance of the HILDA.D T and HILDA.P T variants.

Other approaches that stand out in Figures 5.4 and 5.5 are the HILDA.D H I and HILDA.P H I methods. These methods employ a comparably deep analysis of the full rhetorical structure of a document or its paragraphs, respectively. Such an analysis typically outperforms shallower analyses, e.g., the L and – especially – T analysis methods. However, the combination of a rather naive weighting scheme I with a deep, hierarchy-based analysis of the full rhetorical structure of a document or its paragraphs turns out to result in a rather narrow focus on very specific parts of a document, while effectively ignoring most of the document. This causes the HILDA.D H I and HILDA.P H I methods to be amongst the weakest of our considered methods.

Approaches that stand out in a positive way are those applying the hierarchy-based RST analysis H to sentence-level RST trees generated by HILDA or SPADE, with weighting schemes X and F, i.e., HILDA.S H X, HILDA.S H F, SPADE.S H X, and especially SPADE.S H F. These approaches perform comparably well because they involve a detailed analysis of a text's rhetorical structure in all possible ways – the analysis is performed on the smallest considered units, i.e., sentences, the hierarchy of the full RST trees is taken into account, and the weights are differentiated per type of rhetorical relation.

Relation	Satellite description
Attribution	Clauses containing reporting verbs or cognitive predicates related to reported messages presented in nuclei.
Background	Information helping a reader to sufficiently comprehend matters presented in nuclei.
Contrast	Situations juxtaposed to and compared with situations in nuclei, which are considered as mostly similar, yet different in a few respects.
Elaboration	Additional detail about matters presented in nuclei.
Enablement	Information increasing a reader's potential ability of performing actions presented in nuclei.

Table 5.6: Most common RST relation types in our corpus.

All in all, our work makes three important contributions. First, our novel polarity classification approaches guided by deep leaf-level or hierarchy-based analyses of a text's rhetorical structure significantly outperform our baselines that are not guided by RST, i.e., Baseline and Position, and our shallow RST-based baselines SPADE.S T I and SPADE.S T II. Second, our novel RST-based weighting schemes in which we differentiate the weights of satellites, or both nuclei and satellites, by their RST relation type significantly outperforms existing, more naive weighting schemes. Third, our comparison of the performance of polarity classification approaches guided by sentence-level, paragraph-level, and document-level RST trees reveals that RST-based polarity classification works best when focusing on RST trees of smaller units of a text, e.g., sentences.

5.4.2.6 Optimized Weights

The observed superior performance of our novel weighting schemes X and F originates in their optimized weights that assign distinct rhetorical elements different roles in conveying a document's overall sentiment. These weights have been optimized for each considered RST-based approach separately, for ten folds each. The optimized weights for weighting schemes X and F – shown in Figures 5.6 and 5.7, respectively – exhibit several patterns for the most common RST relations in our corpus, detailed in Table 5.6.

For weighting scheme X, Figure 5.6 shows box plots of the optimized diminishing factors and the optimized weights for the most common RST relation types in our corpus, for RST-based polarity classification methods guided by sentence-level, paragraph-level, and document-level RST trees. In general, the degree of dispersion for optimized weights for nucleus elements is lower than it is for satellites. This indicates that the importance assigned to nuclei is rather consistent across all of our considered methods, whereas the importance of satellite elements appears to be comparably harder to identify.

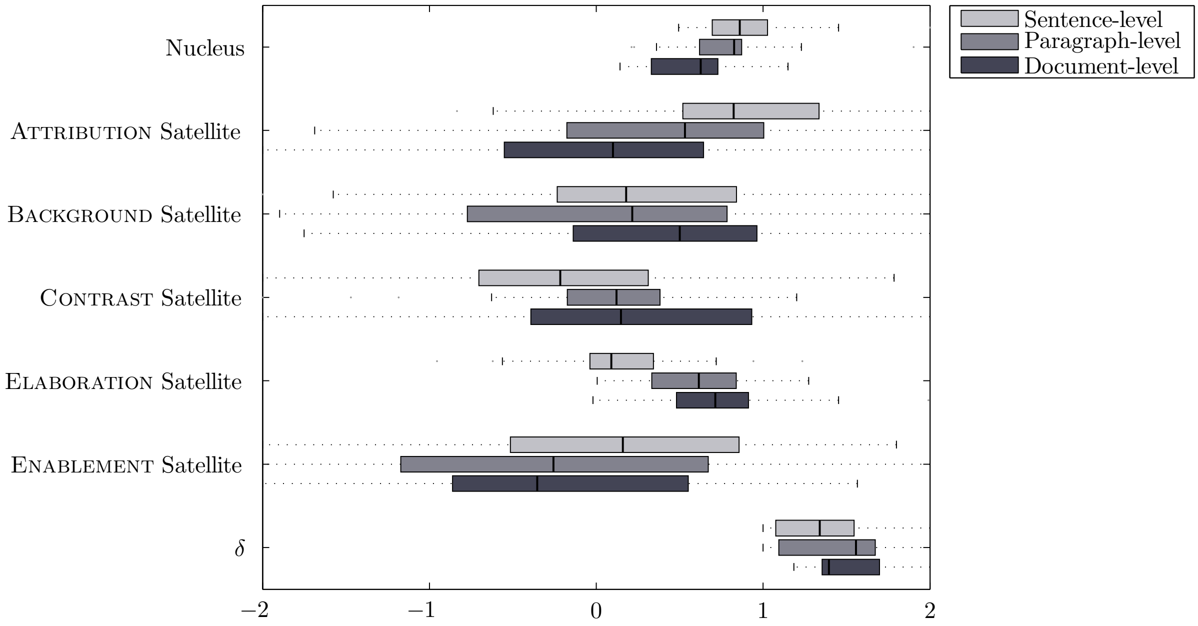


Figure 5.6: Box plots of the optimized values of the most common RST relation types' weights, as well as of the diminishing factor δ for weighting scheme X, for RST-based approaches guided by sentence-level, paragraph-level, and document-level RST trees. The whiskers signal the extreme data points within 1.5 times the interquartile range above and below the third and first quartile, respectively.

Similarly, albeit to a more limited extent, the spread of the optimized weights for document-level RST trees tends to exceed the degree of dispersion of optimized weights for paragraph-level RST trees, which in turn shows a tendency of exceeding the spread of optimized weights for sentence-level RST trees. Apparently, the larger an RST tree, the more difficult it is to find a set of weights that generalizes well. This provides a partial explanation for the observed superiority of our considered sentence-level RST-based approaches over the paragraph-level and document-level methods (see Section 5.4.2.5), in addition to the explanations already discussed in Section 5.4.2.5.

In the optimized weights for weighting scheme X, nucleus elements are generally considered to be relatively important, with most weights ranging between approximately 0.5 and 1. The sentiment expressed in Elaboration satellites tends to be assigned a similar or somewhat lower importance. Satellites that have a Contrast relation with their respective nuclei tend to receive weights around or even below 0. Negative weights are assigned to contrasting satellites especially when guiding the analysis by sentence-level RST trees. Background satellites are, on average, typically assigned relatively low weights as well, albeit with a rather high degree of dispersion, with most of their weights typically ranging from approximately 0 to 1.

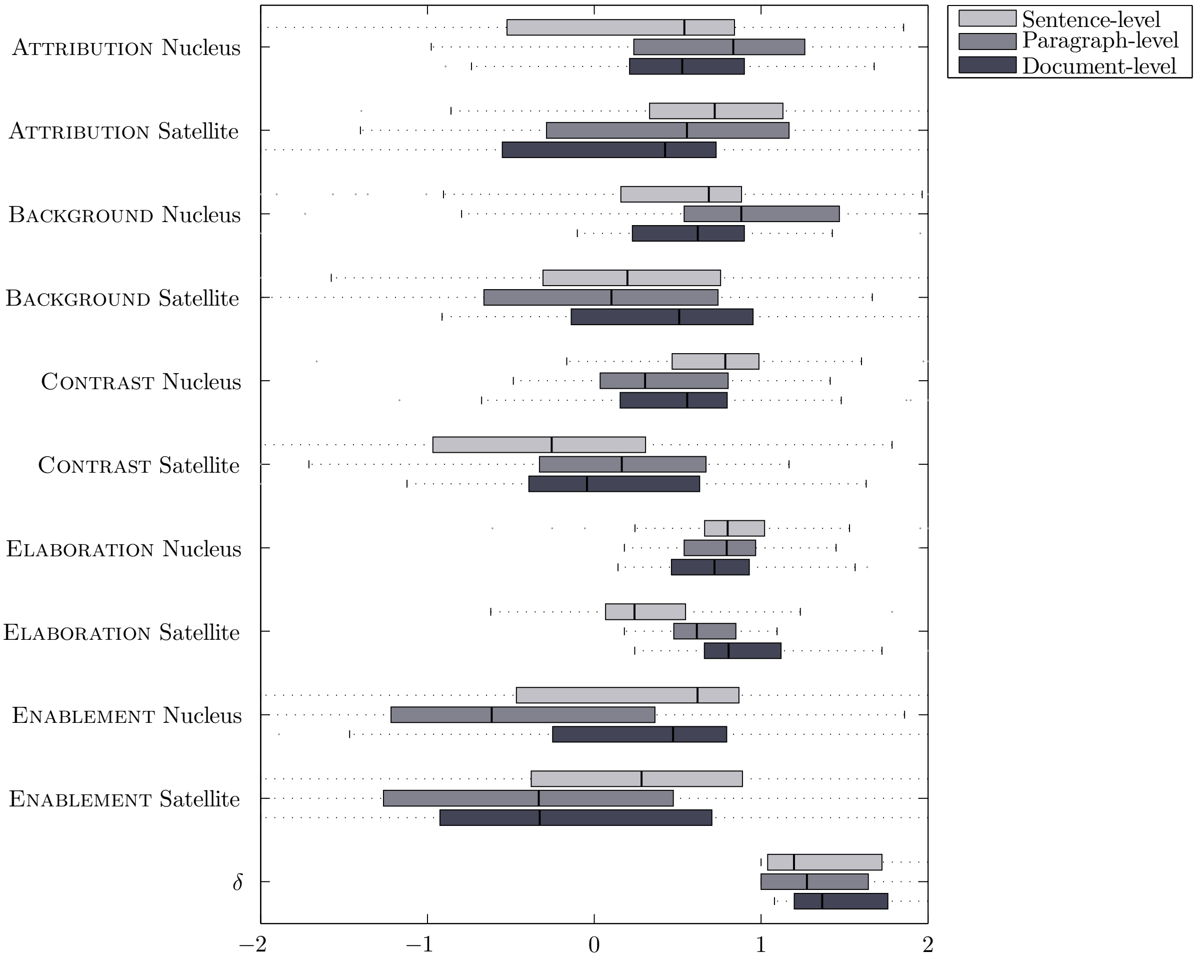


Figure 5.7: Box plots of the optimized values of the most common RST relation types' weights, as well as of the diminishing factor δ for weighting scheme F, for RST-based approaches guided by sentence-level, paragraph-level, and document-level RST trees. The whiskers signal the extreme data points within 1.5 times the interquartile range above and below the third and first quartile, respectively.

No conclusive pattern is exhibited by the weights assigned to Attribution satellites, i.e., the text segments containing reporting verbs or cognitive predicates related to reported messages in nuclei. For our considered collection of Englishs movie reviews, the optimal weights for Attribution satellites appear to range from around 0 for document-level RST trees to around 1 for sentence-level RST trees, with the weights for all levels of analysis showing a rather high degree of dispersion. In addition to the Attribution satellites, the Enablement satellites – i.e., the persuasive segments of the reviews that increase a reader's potential ability of performing actions presented in the nuclei of these reviews – show no conclusive pattern either, as the optimized weights for Enablement satellites range from approximately −1 to 1.

The diminishing factor δ, controlling the weights of the levels of RST trees in the hierarchy-based analyses, is typically optimized to approximately 1.5 for weighting scheme X. In practice, this optimized δ results in the first 15 to 20 levels of an RST tree to have a substantial contribution in the hierarchy-based analysis of the conveyed sentiment. Interestingly, with some (document-level) RST trees being over 100 levels deep in our considered collection of English movie reviews, the optimized diminishing factors effectively mostly disregard the lower, most fine-grained parts of the RST trees. Apparently, there is useful information in the hierarchical rhetorical structure of text, provided that the right balance is found between the level of detail and potential noise in the analysis.

Figure 5.7 shows the optimized values of the diminishing factors and the weights for the most common RST relation types in our corpus, when applying our novel full weighting scheme F to RST-based polarity classification methods guided by sentence-level, paragraph-level, and document-level RST trees. In this weighting scheme, we differentiate the weight of RST elements for both nuclei and satellites by their RST relation type. The weights of the five distinct types of nuclei visualized in Figure 5.7 typically show a somewhat lower degree of dispersion than their associated satellites do. This suggests that, similarly to weighting scheme X, the importance assigned to nuclei is more consistent across all of our considered methods than the importance of satellite elements is. An increasing spread of optimized weights for larger RST trees, as observed for weighting scheme X, is less apparent for weighting scheme F, possibly due to the increased number of degrees of freedom in the optimization process, even for smaller RST trees.

In addition to these general trends, specific patterns can be observed as well for weighting scheme F. First, Attribution nuclei are typically assigned an importance that is rather similar to the importance associated with Attribution satellites. Conversely, for Background and Contrast relations, satellites are more clearly distinct from nucleus elements. Background satellites are typically assigned somewhat less importance than their associated nuclei. Similarly, Contrast satellites are typically assigned lower weights than their associated nuclei, with negative weights not being an exception for the satellites containing information that is contrasting with information presented in nuclei.

For Elaboration and Enablement relations, no conclusive patterns can be observed in the optimized weights for weighting scheme F. For Elaboration relations, the nuclei appear to be more important than the elaborating satellites for sentence-level RST trees, whereas they are more or less equally important for paragraph-level RST trees. Moreover, elaborating satellites are more important than their associated nuclei for document-level RST trees. For the Enablement relation, nuclei and satellites are assigned similar weights in a broad range between approximately -1 and 1.

The optimized values for the diminishing factor δ show a pattern that is comparable to that of weighting scheme X. The typical optimized value for the diminishing factor is around 1.25. With this typical value, the first 30 to 40 levels of the tree are effectively accounted for in the analysis, which entails a deeper analysis than the optimized diminishing factor values for weighting scheme X. Combined with the more detailed set of weights that allows for a distinction between various types of nuclei, this yields a better performance for weighting scheme F, compared to weighting scheme X.

5.4.2.7 Processing a Document

The observed differences in performance of our polarity classification methods originate in how these methods perceive a document in the sentiment analysis process. Interpretations of a document and its conveyed sentiment can be vastly different, depending on the analysis method and weighting scheme used. Figure 5.8 demonstrates how the interpretations of one specific movie review differ across various methods which have exhibited clear differences in polarity classification performance, as detailed in Section 5.4.2.5.

Combined with our utilized sentiment lexicon, our applied tokenization, POS tagging, lemmatization, and word sense disambiguation techniques result in many words in the review being assigned a (somewhat) positive sentiment, whereas fewer words are identified as carrying negative sentiment. The abundance of positive words like "*believable*", "*good*", "*cool*", "*desired*", "*better*", and "*strong*" suggests that the review is rather positive, especially since the review contains only a few negative words, e.g., "*forgotten*", "*bad*", "*silly*", and "*tortured*". However, the review is a negative one, in which the author describes the plot, mentions some good aspects of the movie, and above all stresses several weaknesses of the movie, before recommending the reader not to watch it.

When processing the example review with our absolute baseline approach, i.e., the Baseline method, each part of the review is assigned an equal weight. This results in the document to be interpreted as visualized in Figure 5.8(a). Consequently, the role that some parts of the review have in conveying the overall sentiment does not match their actual role. For instance, the plot details should not be assigned a significant role in conveying the reviewer's negative sentiment, nor should the positive aspects of the movie.

One approach that does a rather rigorous job in focusing on specific parts of the review is the HILDA.P H I approach, i.e., one of our worst performing methods. Figure 5.8(b) demonstrates that the HILDA.P H I method results in an analysis that is highly focused – if not too focused. For this review, the HILDA.P H I method successfully emphasizes the largest part of the reviewer's recommendation and ignores many less relevant parts of the review, such as most of the plot details and the positive aspects of the movie.

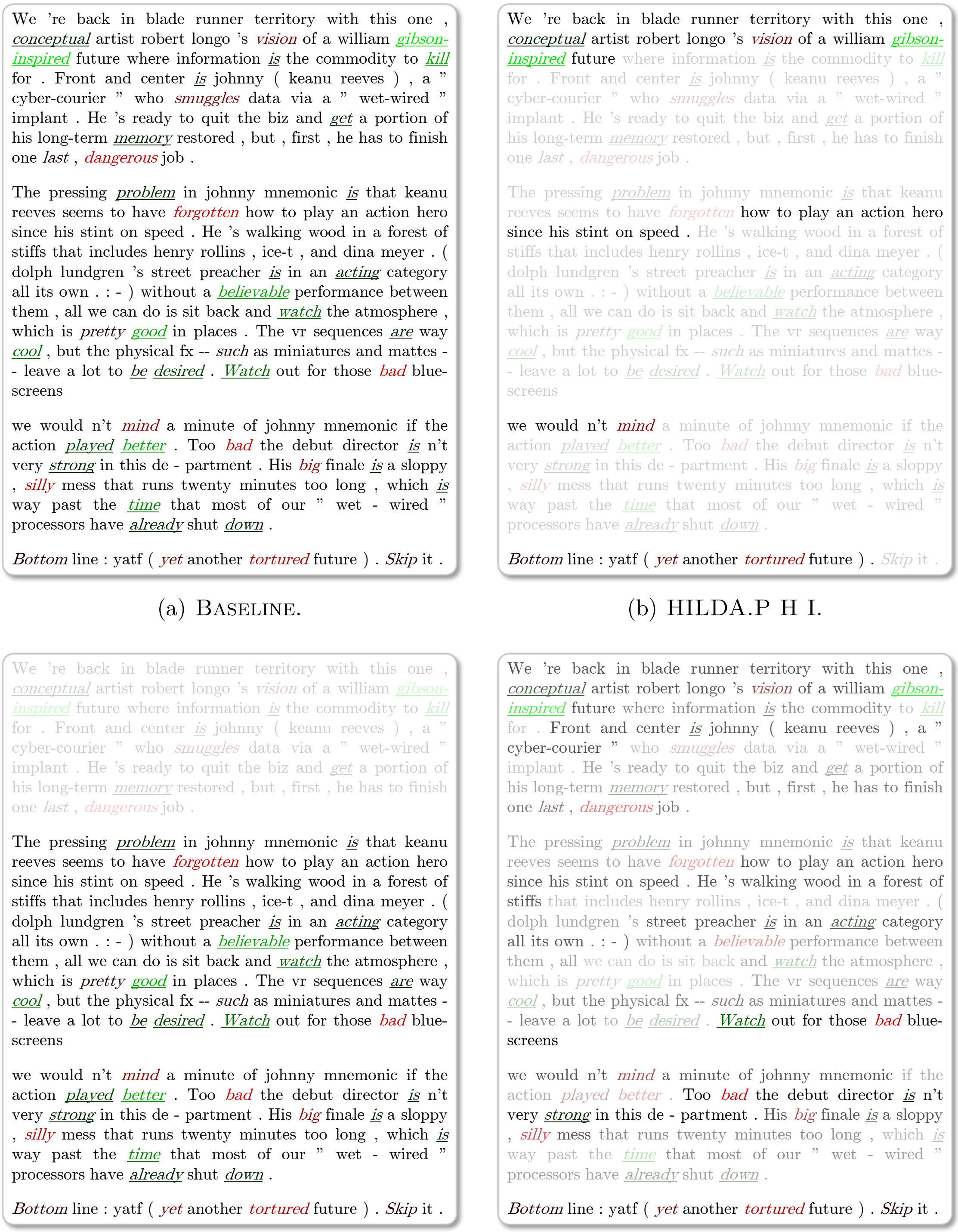
However, the strict selection performed by the HILDA.PHI method additionally causes it to identify only parts of the relevant text segments that support this recommendation, thus leaving it with few cues for the reviewer's sentiment.

A method that focuses considerably less on small, specific parts of our example review is the HILDA.DTX approach, as demonstrated by Figure 5.8(c). Because the RST analysis in this approach is performed only on the top-level splits of the RST trees of the unit of analysis, it is a comparably coarse-grained method. Since its unit of analysis is a document, the HILDA.DTX method effectively differentiates between only two parts of our example review. Its optimized weights cause HILDA.DTX to successfully ignore the first paragraph of the review, which contains irrelevant plot information, and to focus on the remainder of the review instead. However, its nature prevents HILDA.DTX from sufficiently dealing with the nuances in the latter paragraphs.

SPADE.SHF – i.e., our overall best performing approach – yields a very detailed analysis of our example review, by making subtle distinctions between relatively small text segments, as illustrated by Figure 5.8(d). These nuances help bringing out the parts of the review that are most relevant with respect to its overall sentiment. The SPADE.SHF method ignores most of the irrelevant background information in the first paragraph and highlights the reviewer's main concerns in the second and third paragraphs. Moreover, the reviewer's sentiment related to the movie's good aspects is often inverted and mostly ignored. Last, the reviewer's overall recommendation is emphasized in the last paragraph.

5.4.2.8 Caveats

Our experimental results suggest that the best polarity classification performance can be achieved when guiding sentiment analysis by a text's sentential rhetorical structure, rather than by its paragraph-level rhetorical structure, by its document-level rhetorical structure, or by no RST-based aspects of content at all. The superiority of sentence-level RST-based polarity classification over paragraph-level and document-level RST-based analyses may be partly caused by the quality of the RST parsers used for performing sentence-level RST analyses on the one hand, and paragraph-level and document-level RST analyses on the other hand. The SPADE parser yields slightly better results than the HILDA parser, when utilized for analyzing sentential rhetorical structure. This suggests that there is indeed some difference in quality of the RST analyses performed by the parsers. Yet, the superiority of sentence-level RST-based analyses can be observed within the various HILDA-based methods too, thus indicating that this phenomenon is not solely attributable to parser quality, but to the alternative explanations provided in Sections 5.4.2.5 and 5.4.2.6 as well.



We 're back in blade runner territory with this one , *conceptual* artist robert longo 's *vision* of a william *gibson-inspired* future where information *is* the commodity to *kill* for . Front and center *is* johnny (keanu reeves) , a ” cyber-courier ” who *smuggles* data via a ” wet-wired ” implant . He 's ready to quit the biz and *get* a portion of his long-term *memory* restored , but , first , he has to finish one *last* , *dangerous* job .

The pressing *problem* in johnny mnemonic *is* that keanu reeves seems to have *forgotten* how to play an action hero since his stint on speed . He 's walking wood in a forest of stiffs that includes henry rollins , ice-t , and dina meyer . (dolph lundgren 's street preacher *is* in an *acting* category all its own . : -) without a *believable* performance between them , all we can do is sit back and *watch* the atmosphere , which is *pretty* *good* in places . The vr sequences *are* way *cool* , but the physical fx -- *such* as miniatures and mattes - - leave a lot to *be* *desired* . *Watch* out for those *bad* blue-screens

we would n't *mind* a minute of johnny mnemonic if the action *played* *better* . Too *bad* the debut director *is* n't very *strong* in this de - partment . His *big* finale *is* a sloppy , *silly* mess that runs twenty minutes too long , which *is* way past the *time* that most of our ” wet - wired ” processors have *already* shut *down* .

Bottom line : yatf (*yet* another *tortured* future) . *Skip* it .

(a) BASELINE.

We 're back in blade runner territory with this one , *conceptual* artist robert longo 's *vision* of a william *gibson-inspired* future where information *is* the commodity to *kill* for . Front and center *is* johnny (keanu reeves) , a ” cyber-courier ” who *smuggles* data via a ” wet-wired ” implant . He 's ready to quit the biz and *get* a portion of his long-term *memory* restored , but , first , he has to finish one *last* , *dangerous* job .

The pressing *problem* in johnny mnemonic *is* that keanu reeves seems to have *forgotten* how to play an action hero since his stint on speed . He 's walking wood in a forest of stiffs that includes henry rollins , ice-t , and dina meyer . (dolph lundgren 's street preacher *is* in an *acting* category all its own . : -) without a *believable* performance between them , all we can do is sit back and *watch* the atmosphere , which is *pretty* *good* in places . The vr sequences *are* way *cool* , but the physical fx -- *such* as miniatures and mattes - - leave a lot to *be* *desired* . *Watch* out for those *bad* blue-screens

we would n't *mind* a minute of johnny mnemonic if the action *played* *better* . Too *bad* the debut director *is* n't very *strong* in this de - partment . His *big* finale *is* a sloppy , *silly* mess that runs twenty minutes too long , which *is* way past the *time* that most of our ” wet - wired ” processors have *already* shut *down* .

Bottom line : yatf (*yet* another *tortured* future) . *Skip* it .

(b) HILDA.P H I.

We 're back in blade runner territory with this one , *conceptual* artist robert longo 's *vision* of a william *gibson-inspired* future where information *is* the commodity to *kill* for . Front and center *is* johnny (keanu reeves) , a ” cyber-courier ” who *smuggles* data via a ” wet-wired ” implant . He 's ready to quit the biz and *get* a portion of his long-term *memory* restored , but , first , he has to finish one *last* , *dangerous* job .

The pressing *problem* in johnny mnemonic *is* that keanu reeves seems to have *forgotten* how to play an action hero since his stint on speed . He 's walking wood in a forest of stiffs that includes henry rollins , ice-t , and dina meyer . (dolph lundgren 's street preacher *is* in an *acting* category all its own . : -) without a *believable* performance between them , all we can do is sit back and *watch* the atmosphere , which is *pretty* *good* in places . The vr sequences *are* way *cool* , but the physical fx -- *such* as miniatures and mattes - - leave a lot to *be* *desired* . *Watch* out for those *bad* blue-screens

we would n't *mind* a minute of johnny mnemonic if the action *played* *better* . Too *bad* the debut director *is* n't very *strong* in this de - partment . His *big* finale *is* a sloppy , *silly* mess that runs twenty minutes too long , which *is* way past the *time* that most of our ” wet - wired ” processors have *already* shut *down* .

Bottom line : yatf (*yet* another *tortured* future) . *Skip* it .

(c) HILDA.D T X.

We 're back in blade runner territory with this one , *conceptual* artist robert longo 's *vision* of a william *gibson-inspired* future where information *is* the commodity to *kill* for . Front and center *is* johnny (keanu reeves) , a ” cyber-courier ” who *smuggles* data via a ” wet-wired ” implant . He 's ready to quit the biz and *get* a portion of his long-term *memory* restored , but , first , he has to finish one *last* , *dangerous* job .

The pressing *problem* in johnny mnemonic *is* that keanu reeves seems to have *forgotten* how to play an action hero since his stint on speed . He 's walking wood in a forest of stiffs that includes henry rollins , ice-t , and dina meyer . (dolph lundgren 's street preacher *is* in an *acting* category all its own . : -) without a *believable* performance between them , all we can do is sit back and *watch* the atmosphere , which is *pretty* *good* in places . The vr sequences *are* way *cool* , but the physical fx -- *such* as miniatures and mattes - - leave a lot to *be* *desired* . *Watch* out for those *bad* blue-screens

we would n't *mind* a minute of johnny mnemonic if the action *played* *better* . Too *bad* the debut director *is* n't very *strong* in this de - partment . His *big* finale *is* a sloppy , *silly* mess that runs twenty minutes too long , which *is* way past the *time* that most of our ” wet - wired ” processors have *already* shut *down* .

Bottom line : yatf (*yet* another *tortured* future) . *Skip* it .

(d) SPADE.S H F.

Figure 5.8: Movie review cv817_3675, processed by various methods. Negative words are printed in italics, whereas positive words are underlined and printed in italics. Sentiment-carrying words with high intensity are brighter. Text segments that are assigned a relatively low weight in the analysis of the conveyed sentiment are more transparent than text segments with higher weights.

A failure analysis has revealed that some misinterpretations of the intended sentiment of authors are caused by our methods not specifically accounting for, e.g., negation or amplification of word-level sentiment. Additionally, sarcasm and proverbs are not always interpreted correctly. Furthermore, not all sentiment-carrying words are identified as such, as their word sense is not successfully disambiguated or the word cannot be found in our sentiment lexicon.

Other challenges are related to the information content of documents. For instance, some authors of the reviews in our corpus evaluate a movie by comparing it with other movies, of which their judgment is moreover often implicit. These statements are particularly problematic to interpret correctly, as they require the processing of comparisons, a distinction between entities and their associated sentiment, and the incorporation of real-world knowledge into the analysis. Another source of errors lies in the tendency of some reviewers to mix their opinionated statements with plot details that may contain sentiment-carrying words that do not signal the reviewers' personal sentiment per se.

Even though our RST-based polarity classification methods cannot cope particularly well with the specific phenomena mentioned above, they do perform significantly better than our non-RST baselines. It should however be noted that the observed significant improvements in terms of performance come at a cost of increased processing time. Applying a typical RST-based approach increases the required processing time with about a factor 10 with respect to either the BASELINE approach or the POSITION method. The bottleneck here is formed by the SPADE and HILDA parsers, rather than the application of our weighting schemes. As such, this shortcoming could be compensated for in future work by utilizing another, possibly faster RST parser.

Nevertheless, in spite of these caveats, our experimental results clearly demonstrate the potential of RST-based polarity classification. With the right set of weights, significant polarity classification performance improvements over traditional approaches can be achieved by guiding the sentiment analysis process by rhetorical relations or, even better, by the rhetorical *structure* of text, as identified by performing a document-level, paragraph-level, or, preferably, sentence-level RST analysis.

5.5 Conclusions

In this chapter, we have demonstrated that automated sentiment analysis should be guided by a deep and fine-grained analysis of a text's rhetorical structure. Such a linguistic analysis enables the distinction between important text segments and less important ones in terms of their contribution to a text's overall sentiment, based on their rhetorical roles.

This is a significant step forward with respect to existing work, which has been limited to guiding sentiment analysis by shallow analyses of rhetorical relations in (mostly sentence-level) rhetorical structure trees. We have argued that it is crucial to account for a text's rhetorical *structure* rather than such isolated rhetorical *relations* in order to obtain a better understanding of a text's conveyed sentiment.

In this light, we have proposed to harvest information from full rhetorical structure trees in order to realize a more accurate structural analysis of a text and its conveyed sentiment than existing work can. Rather than guiding sentiment analysis by coarse-grained segmentations stemming from the top-level splits in sentence-level rhetorical structure trees, we guide the sentiment analysis process by the rhetorical roles of the leaf nodes of sentence-level, paragraph-level, and document-level rhetorical structure trees and we additionally account for the full rhetorical structure in which these roles are defined. Our experimental results on a corpus of English movie reviews demonstrate the superiority of our endeavors over existing work in terms of polarity classification performance.

All in all, the contribution of this chapter is three-fold. First, our novel polarity classification methods guided by deep leaf-level or hierarchy-based analyses of the rhetorical structure of a text significantly outperform existing approaches that are are guided by shallow RST analyses, or by no RST-based analyses at all. Second, our novel RST-based weighting schemes in which we differentiate the weights of satellites, or both nuclei and satellites, by their RST relation type significantly outperforms existing naive weighting schemes. Third, we have compared the performance of polarity classification approaches guided by sentence-level, paragraph-level, and document-level RST trees, thus revealing that RST-based polarity classification works best when focusing on RST trees of smaller units of a text, such as sentences.

In future work, we aim to validate our findings on other corpora, covering other domains or other types of text. Additional challenges lie in improving the scalability of our method as well as in applying our findings to machine learning approaches to sentiment analysis. Therefore, the next chapter of this dissertation deals with a more scalable method for exploiting (discourse) structure of texts in sentiment analysis. The subsequent chapter investigates the applicability of our findings in a machine learning approach rather than a lexicon-based approach to sentiment analysis in order to assess whether our method can improve such methods as well.

Chapter 6

Large-Scale Polarity Ranking with Sentential Rhetorical Structure*

NATURAL *language processing techniques have become of vital importance for many text mining tasks, including sentiment analysis. Existing sentiment analysis tools often rely on simple occurrences of sentiment-carrying words and ignore structural aspects of content. Some methods, however, analyze sentiment based on the discursive role of text segments, as identified by means of, e.g., the Rhetorical Structure Theory (RST). The merits of such computationally intensive analyses have thus far been assessed in very specific, small-scale scenarios. We investigate the usefulness of RST in a large-scale ranking of individual blog posts in terms of their overall polarity. In order to address the computational complexity of RST-based sentiment analysis, we propose to extract key opinionated sentences from blog posts and to subsequently analyze the discourse only for those sentences. Our experimental results show that our large-scale rankings of blog posts, solely based on the sentences thus analyzed, significantly outperform the rankings produced by existing baselines.*

*This chapter is based on the articles "J. Chenlo, A. Hogenboom, and D. Losada. Sentiment-Based Ranking of Blog Posts using Rhetorical Structure Theory. In *18th International Conference on Applications of Natural Language to Information Systems (NLDB 2013)*, volume 7934 of *Lecture Notes in Computer Science*, pages 13–24. Springer, 2013." and "J. Chenlo, A. Hogenboom, and D. Losada. Rhetorical Structure Theory for Polarity Estimation: An Experimental Study. *Data and Knowledge Engineering*, 94:135–147, 2014."

6.1 Introduction

Natural Language Processing (NLP) techniques have become of vital importance for today's information systems (Metais, 2002), which face the challenge of dealing with an ever-increasing amount of user-generated content that is available through the Web. Recent advances in NLP enable information systems to distill actionable knowledge from the abundance of available textual data. For instance, the state-of-the-art in NLP allows for the detection of important events in news messages (Hogenboom et al., 2013c), as well as for the evaluation of news messages and social media posts in terms of their potential effects on stock prices (Mangassarian and Artail, 2007; Schumaker et al., 2012; Yu et al., 2013) or sales (Ghose and Ipeirotis, 2011; Yu et al., 2012). Additionally, NLP techniques enable information systems to detect the topics that people discuss in social media (Cui et al., 2011; Wang et al., 2012), as well as to identify relevant statements (Scholz and Conrad, 2013a) or people's collective viewpoints on such topics (Zhao et al., 2013).

One of the key functionalities of today's information systems is automated opinion mining or sentiment analysis (Cambria et al., 2013; Feldman, 2013; Liu, 2012; Pang and Lee, 2008), which is typically focused on extracting subjective information from natural language text. A typical sentiment analysis task is the classification of the polarity of documents as, e.g., positive or negative. A class of neutral documents may be considered here as well. Another typical sentiment analysis task is the ranking of documents in terms of their associated degree of positivity or negativity with respect to a topic of interest. In this chapter, we focus on the latter sentiment analysis task.

Many commercial sentiment analysis systems mostly rely on simple occurrences of sentiment-carrying words in documents when analyzing the sentiment conveyed by these documents (Feldman, 2013). Yet, in order for information systems to better understand texts and their conveyed sentiment, other aspects than word frequencies alone need to be accounted for (Feldman, 2013; Heerschop et al., 2011a; Hogenboom et al., 2010b; Scholz and Conrad, 2013b). Sentiment may not so much be conveyed by the sentiment-carrying words that people use per se, but rather by the way in which these words are used. In this light, one of the key open research issues in the field of automated sentiment analysis is the role of textual structure in conveying sentiment (Feldman, 2013). Structural aspects may contain valuable information for various NLP tasks (Lioma et al., 2012). These tasks range from paraphrasing and summarization (Bach et al., 2013; Ibrahim and Elghazaly, 2013) to importance ranking (Hogenboom et al., 2012c), text classification (Nguyen and Shirai, 2013), and sentiment analysis (Devitt and Ahmad, 2007; Heerschop et al., 2011a; Hogenboom et al., 2010b; Pang et al., 2002; Taboada et al., 2008).

An increasingly popular way of accounting for structural aspects of content in automated sentiment analysis is to guide the sentiment analysis process by the rhetorical structure of documents (Chardon et al., 2013; Heerschop et al., 2011a; Kennedy and Inkpen, 2006; Polanyi and Zaenen, 2006; Somasundaran et al., 2009a,b; Taboada et al., 2008; Zhou et al., 2011; Zirn et al., 2011). One way of accomplishing this is by applying the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) in order to distinguish important text segments from less important ones with respect to conveying the author's intended sentiment, based on the rhetorical roles (e.g., explanation or contrast) of these respective segments. Thus far, such methods of using RST in sentiment analysis have only been evaluated in very specific settings, i.e., mostly in small-scale document-level polarity classification tasks. In such tasks, RST has been proven to significantly contribute to the overall polarity classification performance – at a cost of computational complexity (Heerschop et al., 2011a; Taboada et al., 2008).

In light of this limiting computational complexity of nevertheless promising RST-based sentiment analysis approaches, the application of RST-based analyses in large-scale sentiment analysis tasks such as polarity ranking tasks for information retrieval is challenging at best. In this chapter, we aim to identify the rhetorical relations that give good guidance for understanding the sentiment conveyed by documents from the blogosphere (i.e., a large-scale, multi-topic domain). Additionally, we aim to quantify the advantage of exploiting these relations in a large-scale polarity ranking task for blog posts. In order to deal with the computational complexity of RST-based sentiment analysis, we build upon recent advances in extracting key opinionated sentences for polarity estimation in blog posts (Chenlo and Losada, 2011) and analyze the structure of the discourse only for selected passages.

The remainder of this chapter is organized as follows. First, in Section 6.2, we discuss related work on existing large-scale sentiment analysis approaches and how structural aspects of content are typically involved in such methods. Then, in Section 6.3, we propose our novel method of RST-based document-level sentiment analysis in a large-scale scenario. We evaluate our method in Section 6.4. Last, we conclude in Section 6.5.

6.2 Related Work

Today's abundance of user-generated content in, e.g., the blogosphere has inspired research on systems that deal with opinions and sentiment, as finding explicit information on user opinions is challenging (Pang and Lee, 2008). A particular challenge lies in the large scale of today's blogosphere.

6.2.1 Large-Scale Sentiment Analysis

The state-of-the-art in automated sentiment analysis has been reviewed extensively (Cambria et al., 2013; Feldman, 2013; Liu, 2012; Pang and Lee, 2008). Existing methods range from machine learning methods, exploiting patterns in vector representations of text, to lexicon-based methods, accounting for the semantic orientation of individual words by matching these words with a sentiment lexicon, which lists words and their associated sentiment. Many hybrid approaches exist as well.

Large-scale sentiment analysis tasks typically pose unique challenges, as their targeted problem is mostly not just in the extraction of sentiment from a large set of documents, but in the identification of relevant (fragments of) documents as well. Numerous studies have been conducted on how to mine opinions in large-scale application areas like the blogosphere. Specifically, the search for relevant subjective documents (regardless of their polarity) has been studied in great detail (Gerani et al., 2010; He et al., 2008a; Santos et al., 2009). Additionally, several effective and efficient methods of finding the opinionated segments of on-topic blog posts have been proposed by Chenlo and Losada (2011). The latter methods enable the representation of the overall opinion of a blog post by a limited number of selected sentences. Such sentences can be selected by combining basic sentence retrieval methods and polarity evidence, i.e., by focusing on the first or last few sentences of a document, or by focusing on the most subjective and on-topic sentences of a document.

In order to characterize the sentiment conveyed by a (selected segment of a) text, many approaches essentially rely on the occurrence of (sentiment-carrying) words (Feldman, 2013). Other aspects of content can however be accounted for as well, with promising features being related to the structure of natural language text. Early work accounts for the absolute position of text segments in the sentiment analysis process (Pang et al., 2002; Pang and Lee, 2004). The results of these studies indicate that the last sentences of a document could be a good indicator of its overall polarity. Positional information has proven to be useful in large-scale sentiment analysis tasks as well. For example, the proximity of query terms to subjective sentences in a document has been used to detect on-topic opinions (Santos et al., 2009). Similarly, a proximity-based opinion propagation method has been proposed to calculate the aggregated opinion at the position of each query term in a document (Gerani et al., 2010).

Other existing work takes into account the semantic cohesion of a document when analyzing its conveyed sentiment (Devitt and Ahmad, 2007), with limited success. The semantic cohesion of a piece of text is accounted for by Devitt and Ahmad (2007) by guiding the sentiment analysis process by the semantic relations between concepts occurring in the text, while assigning the highest importance to concepts with the highest specificity.

In more successful methods, information on text segments' importance for conveying the overall sentiment of a text is harvested from the text's rhetorical structure rather than semantic relations (Chardon et al., 2013; Heerschop et al., 2011a; Kennedy and Inkpen, 2006; Polanyi and Zaenen, 2006; Somasundaran et al., 2009a,b; Taboada et al., 2008; Zhou et al., 2011; Zirn et al., 2011). The rhetorical structure of text may be detected by means of an RST analysis.

6.2.2 Rhetorical Structure and Sentiment Analysis

A piece of natural language text can be described by characterizing its structure in terms of the rhetorical relations that hold between parts of the text. Such relations (e.g., explanations or a contrasts) are important for text understanding, because they give information about how the parts of a text are related to one another to form a coherent discourse. An analysis of a text's discourse can help improve the understanding of this text.

Discourse analysis is concerned with how meaning is built up in the larger communicative process. Such an analysis can be applied on different levels of abstraction, i.e., within a sentence, within a paragraph, or within a document or conversation. The premise is that each part of a text has a specific role in conveying the overall message.

RST (Mann and Thompson, 1988) is one of the leading discourse theories. The theory can be used to split texts into segments that are rhetorically related to one another. Each segment may in turn be split as well, thus yielding a hierarchical rhetorical structure. Within this structure, text segments can be either nuclei or satellites, with nuclei being assumed to be more significant than satellites with respect to understanding and interpreting a text. Many types of relations between text segments exist. A satellite may for instance be an explanation of what is explained in a nucleus. It can also form a contrast with respect to matters presented in a nucleus.

In order to exploit the rhetorical structure of text in automated sentiment analysis, some existing methods (Heerschop et al., 2011a; Taboada et al., 2008; Zhou et al., 2011; Zirn et al., 2011) rely more strongly on rhetorical relations as defined in RST than others (Chardon et al., 2013; Kennedy and Inkpen, 2006; Polanyi and Zaenen, 2006; Somasundaran et al., 2009a,b). The results reported in literature demonstrate the viability of polarity classification guided by rhetorical relations.

In automated sentiment analysis, rhetorical relations are typically used for distinguishing important text segments from less important ones in terms of their contribution to a text's overall sentiment. Early work has made crude distinctions between nuclei and satellites, by assigning satellites low weights, or no weight at all (Taboada et al., 2008).

Later work successfully differentiates between distinct types of satellites when assigning weights to text segments (Heerschop et al., 2011a). Other work disambiguates the sentiment of ambiguous pieces of text – containing conflicting opinions – in a similar fashion by applying a set of rules based on the identified rhetorical roles of text segments (Zhou et al., 2011). Another existing approach is to use Markov logic in order to integrate polarity scores from different sentiment lexicons with information about relations between neighboring segments of texts (Zirn et al., 2011).

One of the challenges of exploiting rhetorical structure of text in sentiment analysis is the processing time required for identifying discourse structure in natural language text (Heerschop et al., 2011a). This problem seems to thwart the applicability of such methods in large-scale scenarios. However, when we combine RST-based sentiment analysis techniques with effective and efficient methods that enable the representation of the overall opinion of a text by a limited number of selected sentences (Chenlo and Losada, 2011), the state-of-the-art in large-scale sentiment analysis may be significantly improved.

6.3 Large-Scale Sentiment-Based Ranking Guided by Rhetorical Structure

With the blogosphere being one of the most important sources of opinions in social media, recent research efforts have been focused on detecting opinions in blog posts (Santos et al., 2012). Classical information retrieval techniques are not sufficient for building information systems that effectively deal with the opinionated nature of data in the blogosphere (Pang and Lee, 2008). Therefore, large-scale opinion mining is typically approached as a two-stage process that involves an initial topic retrieval stage for retrieving relevant posts given a user query, and a subsequent re-ranking stage that takes into account opinion-based features (Ounis et al., 2008). This second stage can also be subdivided into two different subtasks, i.e., an opinion-finding task, where the main aim is to find opinionated blog posts related to the query, and a subsequent polarity estimation task that aims to identify the orientation of a blog post with respect to the topic (e.g., positive or negative).

The latter polarity estimation task is a challenging task with many unresolved issues, such as irony and conflicting opinions. Yet, typical approaches are rather naive methods that essentially estimate the polarity of a text based on the frequencies of positive and negative terms (Feldman, 2013). In line with recent findings (Feldman, 2013; Heerschop et al., 2011a; Hogenboom et al., 2010b; Scholz and Conrad, 2013b), we argue that the polarity estimation problem cannot be dealt with by using frequency-based techniques alone.

In fact, the findings of Ounis et al. (2008) and Chenlo and Losada (2011) demonstrate that most lexicon-based approaches fail to retrieve more positive or negative documents than baselines that do not account for the polarity of individual terms at all. This phenomenon may be caused by the polarity of a document being not so much conveyed by the sentiment-carrying words that people use, but rather by the way in which these words are used. Rhetorical roles of text segments and their relative importance should therefore be accounted for when determining the overall polarity of a text (Chardon et al., 2013; Heerschop et al., 2011a; Kennedy and Inkpen, 2006; Polanyi and Zaenen, 2006; Somasundaran et al., 2009a,b; Taboada et al., 2008; Zhou et al., 2011; Zirn et al., 2011), for instance by incorporating RST into the sentiment analysis process.

The application of RST in large-scale opinion search tasks is challenging due to the high computational complexity of RST-based sentiment analysis. Moreover, in the blogosphere, noise is introduced by the presence of spam, off-topic information, or relevant yet non-opinionated content. This harms the effectiveness of opinion finding techniques. Therefore, we build upon recent advances in extracting key opinionated sentences for polarity estimation in blog posts (Chenlo and Losada, 2011). We propose to analyze the structure of the discourse only for those selected sentences.

6.3.1 Finding Relevant Polar Sentences

We start with a list of documents, where each document d has been assigned a topic relevance score ρ_{dq} with respect to a query q. We assume the documents to be ordered by decreasing topic relevance. In this ranked list, we search for on-topic opinions by jointly applying an effective and efficient sentence retrieval method (Chenlo and Losada, 2011) and a well-known sentiment classifier, i.e., OpinionFinder (Wilson et al., 2005). Following Chenlo and Losada (2011), we determine the polarity ζ_{sq} of a sentence s with respect to a query q as a linear combination of the normalized relevance $\widetilde{\rho}_{sq}$ of the sentence with respect to the query, and the polarity ζ_s of the sentence, i.e.,

$$\zeta_{sq} = \beta \widetilde{\rho}_{sq} + (1 - \beta)\, \zeta_s, \tag{6.1}$$

with $\beta \in [0, 1]$ being a free parameter.

The normalized relevance $\widetilde{\rho}_{sq}$ of a sentence s with respect to a query q is computed by normalizing the tf-idf$\,(s, q)$ score that quantifies the relatedness between s and q, i.e.,

$$\widetilde{\rho}_{sq} = \frac{\text{tf-idf}\,(s, q)}{\text{tf-idf}\,(s^*, q)}. \tag{6.2}$$

In (6.2), the tf-idf (s^*, q) score associated with the most relevant sentence s^* is computed as the maximum score over all sentences S from the ranked list of documents, i.e.,

$$s^* = \arg\max_{s' \in S} \text{tf-idf}\,(s', q)\,. \tag{6.3}$$

In order to compute the tf-idf (s, q) score of a sentence s with respect to a query q, we use the BM25-like tf-idf implementation of Lemur (Zhai, 2001), as BM25 is a robust and effective information retrieval model that has shown its merits in many search tasks (Robertson, 2005). In Lemur, the tf-idf (s, q) score is computed as

$$\text{tf-idf}\,(s, q) = \sum_{t \in s \cap q} \left(\frac{k_1 f_{ts}}{f_{ts} + k_1 \left(1 - b + b \frac{|s|}{\sum_{s' \in S} \frac{|s'|}{|S|}} \right)} \cdot \frac{1{,}000 f_{tq}}{f_{tq} + 1{,}000} \cdot \log \left(\frac{|S|}{|S_t|} \right)^2 \right), \tag{6.4}$$

where f_{ts} and f_{tq} represent the frequencies of term t in a sentence s and a query q, respectively, $|s|$ and $|s'|$ quantify the respective length (i.e., word count) of sentence s and s' from all sentences S, $|S|$ models the number of sentences in S, and $|S_t|$ is the number of sentences in S containing the term t. Last, k_1 and b are free parameters, for which we apply the recommended values $k_1 = 1.2$ and $b = 0.75$ (Robertson and Walker, 2000).

The second component of the polarity ζ_{sq} of a sentence s with respect to a query q is the polarity ζ_s of sentence s. This polarity score is based on OpinionFinder's output. The information provided by OpinionFinder has been proven useful for both subjectivity and polarity estimation in various experimental studies (Chenlo and Losada, 2011; He et al., 2008a,b; Santos et al., 2009). OpinionFinder exploits semantics in a machine learning method that estimates which sentences in a document are subjective, and additionally marks the opinion holders and the positive and negative words in those sentences.

Following existing work (Chenlo and Losada, 2011), we use the terms tagged by OpinionFinder as positive or negative as indicators for the positive or negative polarity score of a sentence, respectively. To this end, ζ_s represents the number of identified terms of a specific polarity in the sentence s, relative to the total length of the sentence, i.e.,

$$\zeta_s = \frac{|s \cap P|}{|s|}, \tag{6.5}$$

with P being a set of terms with a specific polarity. Thus, when retrieving positive documents, ζ_s captures the percentage of positive terms in the sentence, whereas for negative document retrieval, ζ_s quantifies the percentage of negative terms in the sentence.

The polarity of a document can be determined based on the polarity scores ζ_{sq} of each selected sentence s with respect to a query q, computed as a function of the relevance and polarity scores of these sentences, as defined in (6.1). The aggregated score can be computed as the average score of *all* polar sentences, the average score of the *first* or *last* n polar sentences, or the average score of the n sentences with the *highest* polarity score (Chenlo and Losada, 2011). The latter method, PolMeanBestN, has been shown to be very robust, as well as to be the best performing approach for polarity estimation of blog posts (Chenlo and Losada, 2011).

Therefore, in this chapter, we use PolMeanBestN in order to identify the key evaluative sentences. We adopt the best configuration obtained for PolMeanBestN by Chenlo and Losada (2011), i.e., $n = 1$. This means that for each blog post, we extract only one sentence – i.e., the sentence with the highest polarity with respect to a query – in order to estimate the the blog post's overall polarity with respect to the query. We subject these selected sentences to further analysis in order to better determine the orientation of individual blog posts as a whole.

6.3.2 Parsing Sentential Rhetorical Structure

The sentence-level polarity score ζ_s for a key evaluative sentence s may not be very accurate in case it has been computed by applying (6.5), as other aspects than word frequencies alone need to be accounted for in order to better understand sentences and their conveyed sentiment (Feldman, 2013; Heerschop et al., 2011a; Hogenboom et al., 2010b; Scholz and Conrad, 2013b). We argue that the way of computing ζ_s, as described in (6.5), can be improved by involving discourse analysis into the sentiment analysis process. Sentences can be decomposed into different parts, each of which fulfills a specific rhetorical role and should hence be treated accordingly in the analysis. An evaluation of these so-called discourse units and their interrelations could help us to get a more reliable sentiment score (Heerschop et al., 2011a; Taboada et al., 2008).

In order to identify the discourse structure of our identified key evaluative sentences, we use a tool for Sentence-level PArsing of DiscoursE (SPADE), which was created by Soricut and Marcu (2003). The SPADE parser creates RST trees for individual English sentences. SPADE was trained and tested on the train and test set of the RST Discourse Treebank (RST-DT) (Carlson et al., 2003), achieving an F_1 score of 83.1% on identifying the right rhetorical relations and their correct arguments (Soricut and Marcu, 2003). The discourse relations identified by SPADE that are taken into account in the work reported on in this chapter are detailed in Table 6.1.

Relation	Satellite description
Attribution	Clauses containing reporting verbs or cognitive predicates related to reported messages presented in nuclei.
Background	Information helping to comprehend matters presented in nuclei.
Cause	An event leading to a result presented in the nucleus.
Comparison	Examination of matters along with matters presented in nuclei.
Condition	Hypothetical, future, or otherwise unrealized situations, the realization of which influences the realization of nucleus matters.
Consequence	Information on the effects of events presented in nuclei.
Contrast	Situations juxtaposed to and compared with situations in nuclei, which are mostly similar, yet different in a few respects.
Elaboration	Additional detail about matters presented in nuclei.
Enablement	Information increasing a reader's potential ability of performing actions presented in nuclei.
Evaluation	Evaluative comments about matters presented in nuclei.
Explanation	Justifications or reasons for situations presented in nuclei.
Joint	No specific relation holds with the matters presented in nuclei.
Otherwise	A situation of which the realization is prevented by the realization of the situation presented in the nucleus.
Temporal	Events with an ordering in time with respect to events in nuclei.

Table 6.1: RST relations taken into account in our experiments.

6.3.3 Using the Structure of Key Sentences in Polarity Ranking

We propose to recompute the polarity score ζ_s of each selected key evaluative sentence s by exploiting its discourse structure. In this process, we propose to account for a sentence's structure by assigning distinct weights to specific parts of the sentence, based on the identified rhetorical roles of these parts. Following existing work (Heerschop et al., 2011a; Taboada et al., 2008), we differentiate between rhetorical roles as identified in the first (i.e., top-level) split of the sentence-level RST trees generated by SPADE. As such, we recompute the polarity score ζ_s of a key evaluative sentence s as a weighted sum of the polar terms occurring in its identified segments s_i, i.e.,

$$\zeta_s = \sum_{s_i \in s} \left(w_{s_i} \frac{|s_i \cap P|}{|s_i|} \right), \tag{6.6}$$

where w_{s_i} is the weight for segment s_i, representing (all terms occurring in) a top-level nucleus or satellite in the sentence-level RST tree generated for sentence s by SPADE. P is a set of terms with a specific polarity. Observe that the weights w_{s_i} are free parameters that need to be trained for each distinct rhetorical relation. Additionally, note that the computationally intensive RST analysis can be done off-line, i.e., at indexing time.

Combined with the relevance of each key evaluative sentence s with respect to a query q, the polarity score ζ_s thus computed provides an intricate estimation of the polarity ζ_{sq} of the sentence s with respect to the query q, as defined in (6.1). We propose to use these sentence-level polarity scores in order to estimate the polarity ζ_{dq} of their associated document d with respect to the query q. The documents can then be ranked based on these document-level polarity scores.

The overall document re-ranking process is defined as follows. Given an initial list of documents, ranked by decreasing relevance score, we re-rank the list to promote on-topic blog posts that are positively or negatively opinionated. Documents are re-ranked based on their computed polarity score ζ_{dq}, which we define as

$$\zeta_{dq} = \gamma \widetilde{\rho}_{dq} + (1 - \gamma)\, \zeta^{\mathrm{S}}_{dq}, \tag{6.7}$$

where $\widetilde{\rho}_{dq}$ is the topic relevance score ρ_{dq} of document d with respect to query q after a query-based normalization in the interval $[0, 1]$. As we apply PolMeanBestN with $n = 1$, the sentence-based polarity ζ^{S}_{dq} of document d with respect to query q is defined as

$$\zeta^{\mathrm{S}}_{dq} = \max_{s \in d} \zeta_{sq}, \tag{6.8}$$

with $\gamma \in [0, 1]$ being a free parameter. As a result, every document is re-ranked based on topic relevance and the positivity or negativity of the polar sentence with the highest sentence-level polarity score ζ_{sq}. For β in (6.1) and γ in (6.7), we use values of which Chenlo and Losada (2011) have proven the effectiveness to be very stable across different collections, i.e., $\beta = 0.6$ and $\gamma = 0.6$ for negative polarity estimation, and $\beta = 0.2$ and $\gamma = 0.5$ for positive polarity estimation.

6.4 Evaluation

In order to assess the usefulness of RST-based sentiment analysis in a large-scale scenario, we have evaluated our approach as discussed in Section 6.3 on a large-scale multi-topic data set. The setup of our experiments is detailed in Section 6.4.1, whereas we present our experimental results in Section 6.4.2.

6.4.1 Experimental Setup

Our considered large-scale multi-topic data set is the BLOGS06 document collection of Macdonald and Ounis (2006), which we describe in more detail in Section 6.4.1.1.

The BLOGS06 corpus is one of the most renowned blog test collections, and it contains blog posts with relevance, subjectivity, and polarity assessments. As such, a variety of tasks has been developed for this corpus, and several baselines are available, as discussed in Section 6.4.1.2. In Section 6.4.1.3, we elaborate on how we have used the available data in order to train our models and to assess their performance.

6.4.1.1 Data

The benchmarks of the TREC 2006, TREC 2007, and TREC 2008 blog tracks form the basis of our experiments. All of these tracks have the BLOGS06 corpus as reference corpus. This collection of documents consists of 3,215,171 blog posts, including their comments (if any), which have been crawled from 100,649 blogs over an 11-week period in late 2005 and early 2006. For TREC 2006, TREC 2007, and TREC 2008, collections of 50 query topics were provided, thus resulting in a total of 150 available topics of queries that can be performed on the BLOGS06 corpus. Each query topic is represented by three different fields, i.e., a title, a description, and a narrative.

For each query topic, human judgments were assigned to a pool of relevant documents initially retrieved from the BLOGS06 corpus by the TREC participants. Documents were judged by TREC assessors in two different aspects. First, documents were judged on topic relevance. A document could be relevant, not relevant, or not judged. Second, on-topic documents were judged on their explicit expression of opinion or sentiment about the topic, i.e., no sentiment, positive sentiment, negative sentiment, or mixed sentiment, with the latter category covering not only mixed sentiment, but ambiguous and unclear sentiment as well.

6.4.1.2 Task and Baselines

As of TREC 2008, four distinct tasks can be performed on the BLOGS06 collection for each TREC query topic. One task is the blog distillation task, which focuses on finding blogs with a principal, recurring interest in a query topic. Another task is a baseline ad-hoc blog post retrieval task, which aims to find blog posts about a specific topic, and to assign topic relevance scores to these documents. A third TREC task is the opinion-finding task, where participants are required to identify what people think about a query topic. Last, a polarity task focuses on finding positive or negative opinionated blog posts about a query topic of interest.

In this chapter, we focus on the latter polarity task. An assumption of our model as presented in Section 6.3 is that the topic relevance scores of posts are known a priori.

This is in line with the set-up of the polarity task defined for TREC 2008, where five distinct topic relevance baselines are provided. These baselines have been selected by TREC from the runs submitted to the ad-hoc blog post retrieval task and have varying retrieval effectiveness. The goal is to improve the document rankings of these topic relevance baselines by means of opinion-finding techniques, such that the documents are ranked in accordance with their associated positive or negative sentiment, as well as their topic relevance. Interestingly, many opinion-finding techniques assessed in TREC 2008 were unsuccessful and did not yield better document rankings than the baseline rankings (Ounis et al., 2008). This renders the polarity ranking task a particularly challenging one.

We have used our proposed RST-guided method in order to make a polarity-based ranking of the subset of subjective blog posts retrieved by the standard baseline runs. By doing so, we have omitted those blog posts retrieved by the baselines that have been identified as spam, off-topic, or non-opinionated, as such filtering subtasks are out of our current scope. We have evaluated our performance in terms of the mean average precision (MAP) and the precision at the first ten documents (P@10), as these measures are commonly applied in information retrieval in order to assess ranking performance. We have assessed the statistical significance of observed performance differences by using a paired, two-tailed t-test at the 95% significance level.

6.4.1.3 Training and Testing

In order to be able to assess the performance of our method in the TREC 2008 polarity ranking task, we have preprocessed the 150 available TREC query topics, as well as the subjective blog posts retrieved for these topics by the standard baselines. We have represented each query topic by means of its title only, as the TREC blog track literature (Ounis et al., 2008) demonstrates that titles are succinct query topic descriptions that moreover form the best representation of real users' Web queries. Additionally, we have preprocessed our query representations thus obtained, as well as the blog posts in the BLOGS06 collection, by stemming their text with the Krovetz stemmer (Krovetz, 1993) and by removing 733 English stopwords.

We have constructed a training set and a test set from our preprocessed data, with the training set covering the TREC 2006 and TREC 2007 topics, and the test set covering the TREC 2008 topics. The training set was used for optimizing the parameters of our methods, i.e., the weights assigned to distinct rhetorical roles, whereas the test set was used for assessing the performance of the optimized parameters on unseen data. We have run two separate training and testing procedures focused on maximizing the MAP score, i.e., one for positive polarity ranking and another for negative polarity ranking.

In our training process, we fixed the weight of nuclei to 1 in order to reflect the alleged importance of these elements. Conversely, we assumed the weights of satellites to be real numbers in the interval $[-2, 2]$, thus allowing satellites to contribute positively or negatively to the overall sentiment, as well as to be more important or less important than nuclei. In order to train the parameters of our models, we have used a particle swarm optimization method that has been shown to be an effective method for automatically tuning parameters in information retrieval problems (Parapar et al., 2012).

In our employed particle swarm optimization algorithm, particles search a solution space, where the coordinates of their position correspond with the parameters to be optimized. The fitness of a set of coordinates is modeled as the MAP on our training set. Each iteration, particles move in each dimension with a velocity that is a function of the particle's inertia (i.e., its tendency to move in a straight line at a constant velocity), an increment towards the particle's best known position, and an increment towards the global best position. Following Parapar et al. (2012), we iterate over 100 generations of 25 particles to train our parameters, with inertia and particle increment set to 0.8 and global increment set to 0.95.

6.4.2 Experimental Results

Table 6.2 shows the results of our considered polarity ranking approaches, i.e., the baselines, the baselines enriched with PolMeanBestN, and the baselines enriched with our novel method combining PolMeanBestN with RST-guided sentiment analysis. Each run is evaluated in terms of its ability to retrieve positive and negative documents higher up in the ranking. Our results exhibit several patterns.

6.4.2.1 Polarity Retrieval Performance

Previous findings of Chenlo and Losada (2011) indicate that the PolMeanBestN approach exhibits a performance on all blog posts – including those blog posts that are spam, off-topic, or non-opinionated – that is comparable with the performance of the KLE system of Lee et al. (2008), i.e., the best performing approach at the TREC 2008 blog track (Santos et al., 2012). The results of our current experiments on the subset of subjective blog posts confirm these earlier findings. Table 6.2 demonstrates that the PolMeanBestN method tends to yield small improvements in performance over the five topic relevance baselines provided by TREC, thus rendering it a competitive alternative to the top-ranking methods in the TREC 2008 blog track (Chenlo and Losada, 2011; Santos et al., 2012).

	Positive		Negative	
Method	MAP	P@10	MAP	P@10
Baseline 1	0.266	0.368	0.240	0.296
+PolMeanBestN	0.270	0.372	0.241	0.300
+PolMeanBestN(RST)	**0.273**	**0.374**$^{\triangle\blacktriangle}$	**0.252**	**0.318**$^{\triangle\blacktriangle}$
Baseline 2	0.239	0.334	0.217	0.278
+PolMeanBestN	0.236	0.316	0.222	0.282
+PolMeanBestN(RST)	**0.242**$^{\triangle}$	**0.356**$^{\triangle\blacktriangle}$	**0.226** $^{\blacktriangle}$	**0.310**$^{\triangle\blacktriangle}$
Baseline 3	0.276	**0.350**	0.249	**0.284**
+PolMeanBestN	0.276	0.342	0.252	0.276
+PolMeanBestN(RST)	**0.277**$^{\triangle}$	0.338 $_{\blacktriangledown}$	**0.258**$^{\triangle\blacktriangle}$	0.282
Baseline 4	**0.273**	0.358	0.264	0.274
+PolMeanBestN	0.271	0.350	0.273	0.284
+PolMeanBestN(RST)	0.272	**0.362**$^{\triangle\blacktriangle}$	**0.283**$^{\triangle}$	**0.324**$^{\triangle\blacktriangle}$
Baseline 5	0.239	0.360	0.224	0.300
+PolMeanBestN	0.240	0.358	0.228	0.312
+PolMeanBestN(RST)	**0.279**$^{\triangle\blacktriangle}$	**0.438**$^{\triangle\blacktriangle}$	**0.239**	**0.342**$^{\triangle\blacktriangle}$

Table 6.2: Performance of the TREC baselines for ranking positive and negative blog posts, without sentiment-based information, guided by PolMeanBestN, and guided by our novel RST-based variant of PolMeanBestN. The symbols $^{\triangle}$ ($_{\triangledown}$) and $^{\blacktriangle}$ ($_{\blacktriangledown}$) indicate significant improvements (decreases) over the PolMeanBestN method and the TREC baselines, respectively. The best value in each column for each baseline is printed in bold.

As the PolMeanBestN approach estimates the overall polarity of a blog post by only considering the on-topic sentence in the blog post that has the highest frequency-based polarity score, our findings suggest that a rather straightforward analysis of the sentiment conveyed by the sentence selected by PolMeanBestN from a blog post can already provide a rather good indication for the post's overall polarity. Yet, the results in Table 6.2 additionally indicate that performing a more careful, RST-guided sentiment analysis on a sentence selected by the PolMeanBestN method can yield an even better performance. Our novel method that combines PolMeanBestN with RST-guided sentiment analysis techniques is our overall best performing approach, typically showing significant improvements with respect to both the baselines and PolMeanBestN.

Another observation that can be made from Table 6.2 is that the performance of all approaches on positive document rankings is better than their performance on negative document rankings. This may be caused by negative documents being harder to find in our collection of blog posts, as it contains about 20% more positive posts than negative ones. Additionally, our lexicon-based identification of negative documents may be thwarted by people having a tendency of using rather positive words in order to express negative opinions (Heerschop et al., 2011a).

	Positive		Negative	
Relation	Occurrence	Satellite weight	Occurrence	Satellite weight
Attribution	0.183	0.531	0.177	2.000
Background	0.034	-0.219	0.038	-2.000
Cause	0.009	1.218	0.009	-0.011
Comparison	0.003	-1.219	0.003	-2.000
Condition	0.029	-0.886	0.025	-2.000
Consequence	0.001	0.846	0.001	1.530
Contrast	0.016	-1.232	0.017	-2.000
Elaboration	0.207	2.000	0.219	2.000
Enablement	0.038	2.000	0.038	1.221
Evaluation	0.001	0.939	0.001	-2.000
Explanation	0.007	2.000	0.008	2.000
Joint	0.009	-1.583	0.010	1.880
Otherwise	0.001	-1.494	0.001	-0.428
Temporal	0.003	-2.000	0.003	-0.448

Table 6.3: Optimized satellite weights for RST relation types for positive and negative rankings, along with the occurrence rate of these RST relation types in the training data. The weight of the nuclei for all RST relations was set to 1.

6.4.2.2 Weights for Rhetorical Relations

Table 6.3 reports the weights learned for distinct RST satellite types. Having been assigned a weight of 1 by default, nuclei are assumed to play a comparably important role in conveying the overall sentiment of a piece of text. Yet, the optimized satellite weights reported in Table 6.3 suggest that some types of satellites play an important role as well in conveying a text's overall sentiment. The most meaningful weights are those for satellites of the most frequent RST relations in our training data, i.e., for the Elaboration, Attribution, Enablement, Background, Condition, and Contrast relations.

For both positive and negative documents, satellites elaborating on matters presented in nuclei have typically been assigned relatively high weights, exceeding those assigned to nuclei. Bloggers may thus tend to express their sentiment in a more apparent fashion in text segments that are Elaboration satellites, rather than in the core of the text itself. A similar pattern emerges for persuasive text segments, i.e., Enablement satellites.

Interestingly, text segments in Attribution relations exhibit another pattern, with attributing satellites being more important in negative documents than in positive documents. This suggests that, in our corpus, negative sentiment is conveyed not by a reported message per se, but largely by its reporting clause. Conversely, positive sentiment tends to be largely expressed by the Attribution nucleus, encompassing the reported message. As such, bloggers may be expressing negative sentiment in a comparably indirect way.

The satellites of several other RST relations contribute negatively to the sentiment-based relevance scores of blog posts in our corpus. First, the information in Background satellites has been assigned a negative weight. An explanation may lie in the nature of these satellites, which contain possibly off-topic, and likely less relevant information with respect to a query. Additionally, Condition satellites contribute negatively to sentiment-based relevance scores, as these satellites describe unrealized and hence irrelevant situations. This renders information in Condition satellites perpendicular to a blogger's actual stance on a topic of interest. A similar pattern can be observed for Contrast satellites. These segments contain information that contrasts with the information in the core of the text, and is therefore treated as such by means of our optimized weights.

A caveat with respect to these findings is that some of the optimized weights reported in Table 6.3 tend towards the extreme values used as constraints in the optimization process. This phenomenon is likely to be caused by the sparsity of the considered RST relation types in our corpus, which limits the impact of individual weights for such sparse RST relation types on the classification performance on the corpus as a whole. This renders the optimized weights as such not particularly meaningful per se. Nevertheless, the general patterns exhibited by the weights, as discussed above, can be validated in our considered collection of blog posts.

6.5 Conclusions

In this chapter, we have investigated the usefulness of computationally intensive RST-based analyses in a large-scale sentiment analysis task. We have demonstrated how to successfully perform a large-scale ranking of individual blog posts in terms of their overall document-level polarity by exploiting the rhetorical structure of only a small selection of key evaluative sentences. We have proposed to apply computationally intensive analyses of rhetorical structure to these selected sentences only, and to subsequently assess the sentiment conveyed by these sentences, while accounting for the rhetorical roles of opinionated segments. By doing so, we significantly improve on existing baselines.

The reason for this success lies in our optimized weights for the sentiment conveyed by text segments with distinct rhetorical roles. For instance, our optimized weights account for bloggers' tendency to express their sentiment in a more apparent fashion in elaborating, persuasive, and – to a lesser extent – attributing text segments, rather than in the core segments of the selected sentences. Additionally, the sentiment conveyed by text segments containing off-topic, irrelevant, or contrasting information is typically considered to have a negative contribution to sentiment-based relevance scores of our blog posts.

Our main contribution is that we have taken a first step towards improving the scalability of RST-based sentiment analysis, by guiding the sentiment analysis process by a shallow analysis of the sentential rhetorical structure of only a small fraction of all text. Our results show that an improved understanding of a carefully selected fraction of a text can already yield a significantly better understanding of the text as a whole, in terms of its conveyed sentiment. Our findings thus indicate that a focused RST-based sentiment analysis process can mitigate concerns with respect to the computational complexity of RST analyses, while still enjoying the benefits of a better understanding of a text.

Our results demonstrate the potential of RST-guided sentiment analysis on (selected) sentences. This potential could be further exploited in future work. One possible direction of future work would be to improve the quality of the RST analysis, e.g., by employing a more detailed analysis of the rhetorical structure of relevant parts of a text, or by using more refined representations of rhetorical relations, for instance by applying language models (Lioma et al., 2012). Another direction for future work could be to further improve the scalability of our approach, by exploring more efficient methods for identifying the discourse structure of natural language text. Last, the added value of RST-based information in a machine learning approach rather than a lexicon-based approach to sentiment analysis could be investigated. The latter idea is addressed by the next chapter of this dissertation.

Chapter 7

Polarity Classification Using Structure-Based Vector Representations of Text[*]

The exploitation of structural aspects of content is becoming increasingly popular in rule-based polarity classification systems. Such systems typically weight the sentiment conveyed by text segments in accordance with these segments' roles in the structure of a text, as identified by deep linguistic processing. Conversely, state-of-the-art machine learning polarity classifiers typically exploit patterns in vector representations of texts, mostly covering the occurrence of words or word groups in these texts. However, since structural aspects of content have been shown to contain valuable information as well, we propose to use structure-based features in vector representations of text. We evaluate the usefulness of our novel features on collections of English reviews in various domains. Our experimental results suggest that, even though word-based features are indispensable to good polarity classifiers, structure-based sentiment information provides valuable additional guidance that can help improve the polarity classification performance of machine learning classifiers significantly. The most informative features capture the sentiment conveyed by specific rhetorical elements that constitute a text's core or provide crucial contextual information.

[*] This chapter is based on the article "A. Hogenboom, F. Frasincar, F. de Jong, and U. Kaymak. Polarity Classification Using Structure-Based Vector Representations of Text. *Decision Support Systems*, 74(1):46–56, 2015."

7.1 Introduction

In the past decade, the Web has experienced an exponential growth into a network of more than 555 million Web sites, with over two billion users worldwide (Pingdom, 2012). The Web has become an influential source of information with an increasing share of user-generated content, produced by many contributors (Mangnoesing et al., 2012). This ubiquitous and ever-expanding user-generated content has taken many different forms, including forum messages, (micro)blog posts, and reviews.

The abundance of user-generated content has the potential to act as a catalyst for well-informed decision making, as the data can be used to monitor the wants, the needs, and the opinions of large quantities of (potential) stakeholders, such as customers. Monitoring user-generated content enables decision makers to identify issues and patterns that matter, and to track and predict emerging events (Hogenboom et al., 2014b). However, in this era of Big Data, potentially valuable data is often unstructured, scattered across the Web, and expanding at a fast rate, thus rendering manual analysis of all available data unfeasible (Madden, 2012). Yet, automated tools for information monitoring and extraction can provide timely and effective support for decision making processes.

Today's automated information monitoring and extraction tools can process information from many heterogeneous sources in dynamic environments (Chan, 2006; Chang et al., 2003, 2006) in order to, e.g., detect trending topics in (on-line) conversations (Cui et al., 2011; Wang et al., 2012), or to identify the discussed entities (e.g., products or brands) and the events in which these entities play a role (Hogenboom et al., 2013c). The past decade has brought forth a surge of research interest in extracting one type of valuable information from text in particular – people's sentiment with respect to entities or topics of interest (Balahur et al., 2012; Feldman, 2013; Montoyo et al., 2012; Reyes and Rosso, 2012). A driving force behind this development lies in the significant electronic word-of-mouth effects of subjective user-generated content (Jansen et al., 2009) on, e.g., sales (Rui et al., 2013; Yu et al., 2012) and stock ratings (Yu et al., 2013).

Many automated sentiment analysis techniques are focused on determining the polarity of natural language text, typically by making use of specific cues, e.g., words, parts of words, or other (latent) features of natural language text. This is often done in machine learning approaches (Liu, 2012; Pang and Lee, 2008). However, rule-based sentiment analysis approaches – often relying on sentiment lexicons that list words and their associated sentiment – are attractive alternatives, as the nature of typical rule-based sentiment analysis methods allows for intuitive ways of incorporating deep linguistic analysis into the sentiment analysis process (Heerschop et al., 2011a; Hogenboom et al., 2010b, 2015b).

Solely focusing on explicit cues for sentiment, e.g., words, has been shown not to yield a competitive polarity classification performance (Hogenboom et al., 2014a). Therefore, successful rule-based approaches additionally account for semantic (Hogenboom et al., 2014b) and structural (Chenlo et al., 2013; Devitt and Ahmad, 2007; Heerschop et al., 2011a; Taboada et al., 2008) aspects of content in order to improve polarity classification performance. Such methods typically use a text's structure in order to distinguish important text segments from less important ones in terms of their contribution to the text's overall sentiment, and subsequently weight each segment's conveyed sentiment in accordance with its identified importance.

The performance of competitive rule-based approaches, albeit comparably robust across domains and texts, is typically inferior to the performance of machine learning polarity classification systems (Taboada et al., 2011). The latter systems typically exploit patterns in vector representations of texts, mainly signaling the presence of specific words or word groups in these texts. However, as structural aspects of content have been proven useful in rule-based approaches (Chenlo et al., 2013; Devitt and Ahmad, 2007; Heerschop et al., 2011a; Hogenboom et al., 2015b; Taboada et al., 2008), we propose to incorporate structure-based features in vector representations of text in order to further improve the polarity classification performance of machine learning sentiment analysis methods.

The main contribution of our work lies in our novel structure-based features, which facilitate a richer representation of natural language text that should enable a more accurate classification of its polarity. We evaluate the usefulness of our structure-based features in a machine learning sentiment analysis method. We thus aim to provide insight in the importance of accounting for structural aspects of text in a machine learning approach to sentiment analysis, such that automated sentiment analysis systems can be used more effectively for supporting decision making processes.

The remainder of this chapter is structured as follows. First, in Section 7.2, we provide an introduction to the field of sentiment analysis, with a specific focus on typical features used to represent text in related work, as well as on structure-based sentiment analysis. Then, in Section 7.3, we propose novel, structure-based features that can be used for sentiment analysis. We evaluate the usefulness of our structure-based features for machine learning polarity classification of text in Section 7.4 and we conclude in Section 7.5.

7.2 Related Work

The field of automated sentiment analysis is an upcoming field that has been attracting more and more research initiatives in the past decade (Liu, 2012; Pang and Lee, 2008).

This surge in research interest in automated sentiment analysis techniques is fueled by the potential of sentiment analysis for real-life decision support systems (Cambria et al., 2013; Feldman, 2013). Several trends can be observed in existing sentiment analysis methods, as briefly addressed in Section 7.2.1. The vector representations of text, used by the (performance-wise) most competitive approaches are discussed in Section 7.2.2. In Section 7.2.3, we then elaborate on promising recent advances in sentiment analysis, where the analysis of the sentiment conveyed by a piece of natural language text is guided by the text's structure.

7.2.1 Sentiment Analysis

Existing methods for sentiment analysis focus on various tasks. Some methods deal with distinguishing subjective text segments from objective ones (Wiebe et al., 2004), whereas other approaches have been designed to determine the polarity of words, sentences, text segments, or documents (Pang and Lee, 2008). The latter sentiment analysis task is commonly treated as a binary classification problem, which involves classifying the polarity of a piece of text as either positive or negative. More polarity classes – e.g., classes of neutral or mixed polarity, or star ratings ranging from one to five stars – may be considered as well, yet in this chapter, we address the binary classification problem for the polarity of documents. Existing binary polarity classification approaches range from rule-based to machine learning methods.

Rule-based methods are rather intuitive methods that typically rely on sentiment lexicons, which list explicit sentiment cues like words (Baccianella et al., 2010) or emoticons (Hogenboom et al., 2013a), along with their associated sentiment scores. The scores of individual explicit cues are typically retrieved from a sentiment lexicon and combined in accordance with predefined rules and assumptions, for instance by summing or averaging these scores in order to obtain an overall sentiment score for a text. This overall score is then used as a proxy for the text's polarity class. In the scoring process, negation (Heerschop et al., 2011c; Hogenboom et al., 2011a) or intensification (Taboada et al., 2011) of the sentiment conveyed by specific cues may be accounted for. Moreover, rule-based sentiment analysis allows for intuitive ways of incorporating deep linguistic analysis into the process, for instance by weighting text segments in accordance with their importance, as identified based on their respective rhetorical roles (Heerschop et al., 2011a; Hogenboom et al., 2010b, 2015b). The performance of rule-based methods tends to be comparably robust across domains and texts (Taboada et al., 2011), and the nature of these methods allows for insight into the motivation for assigning a particular polarity class to a text.

Machine learning approaches to polarity classification typically involve building Support Vector Machine (SVM) classifiers or the like, trained for specific corpora by means of supervised methods that aim to exploit patterns in vector representations of natural language text (Taboada et al., 2011). Such classifiers tend to yield comparably high polarity classification accuracy on the collections of texts they have been optimized for (Chaovalit and Zhou, 2005; Kennedy and Inkpen, 2006; Liu, 2012; Pang and Lee, 2008; Taboada et al., 2011), but they require a lot of (annotated) training data, as well as training time in order to reach this performance level. Nevertheless, their superior performance renders machine learning polarity classifiers particularly useful for specific, rather than generic, domain-independent, or corpus-independent applications.

7.2.2 Vector Representations of Text for Sentiment Analysis

Various types of features have been used by existing machine learning approaches to sentiment analysis in order to construct vector representations of text. The most common and most useful features indicate the presence or frequencies of specific single words (i.e., unigrams) or groups of words (i.e., n-grams) (Liu, 2012; Pang and Lee, 2008). Such features constitute a so-called *bag-of-words* vector representation of a text, which in itself has been shown to be rather effective in polarity classification (Pang et al., 2002; Pang and Lee, 2004). Binary features, indicating the presence of specific words, have been shown to outperform features indicating the frequencies of occurrence of words (Pang et al., 2002). Pang and Lee (2008) have suggested that this may indicate that polarity classification differs from (topic-based) text categorization in general in that a text's topic tends to be emphasized by frequent occurrences of certain words, whereas a text's sentiment may not usually be highlighted through repeated use of the same terms. Nevertheless, frequency-based features have been shown to be useful in later work (Paltoglou and Thelwall, 2010).

Another type of information incorporated in typical vector representations of text for sentiment analysis is part-of-speech (POS) information, enabling the distinction between (types of) nouns, verbs, adjectives, and adverbs. Pang and Lee (2008) have argued that the observed correlation between the subjectivity of a piece of text and the presence of adjectives in this text (Hatzivassiloglou and Wiebe, 2000) has been mistakenly taken as evidence of adjectives being good indicators for sentiment, resulting in a possibly misplaced focus on using adjectives as features in the sentiment analysis process (Mullen and Collier, 2004; Turney, 2002; Whitelaw et al., 2005). Other POS types may contribute to sentiment expression too (Pang and Lee, 2008). As such, a more fruitful approach is to differentiate words in the *bag-of-words* representation of a text by their POS (Liu, 2012).

As subjectivity is associated with word meanings rather than lexical representations of words (Bal et al., 2011; Hogenboom et al., 2012a; Mihalcea et al., 2007), it is important to account for semantics when performing sentiment analysis (Hogenboom et al., 2014b). POS information can be useful here to a limited extent (Wilks and Stevenson, 1998), yet more advanced methods involve accounting for semantics by grouping words with similar meanings (Maas et al., 2011; Whitelaw et al., 2005).

Opinion-conveying texts are significantly different from objective texts in terms of occurrences of sentiment-carrying words (van der Meer et al., 2011). In this light, specific sentiment-carrying words have been used as features in so-called *bag-of-sentiwords* vector representations of text, capturing the presence of sentiment-carrying words derived from a sentiment lexicon (Hogenboom et al., 2012b, 2014a). In other work, text has been represented as a *bag-of-opinions*, where features denote occurrences of unique combinations of opinion-conveying words, amplifiers, and negators (Qu et al., 2010). Sentiment scores of text segments have been used as features as well (Hogenboom et al., 2014a). Other features that have been used in vector representations of text for sentiment analysis include features that capture the length of a text segment, and the extent to which it conveys opinions (Mangnoesing et al., 2012).

7.2.3 Structure-Based Sentiment Analysis

Features that capture structural aspects of content have yet to be proposed. Yet, deep linguistic analysis can help dealing with how the semantic orientation of text is determined by the combined semantic orientations of its constituent phrases (Socher et al., 2013). This compositionality can be captured by accounting for the cohesion (Devitt and Ahmad, 2007) or discursive structure (Chardon et al., 2013; Chenlo et al., 2013; Heerschop et al., 2011a; Hogenboom et al., 2015b; Polanyi and Zaenen, 2006; Somasundaran et al., 2009b; Taboada et al., 2008; Zirn et al., 2011) of text in the sentiment analysis process. Such structure-based sentiment analysis methods typically use a text's structure in order to distinguish important text segments from less important ones and subsequently weight each segment's conveyed sentiment in accordance with its assigned importance.

Recent advances in rule-based sentiment analysis suggest that a text's rhetorical structure, as identified by applying the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), can be used for improving polarity classification performance (Chenlo et al., 2013; Heerschop et al., 2011a; Hogenboom et al., 2015b; Taboada et al., 2008). RST is a popular framework for discourse analysis. The RST framework can be used to split a piece of natural language text into segments that are rhetorically related to one another.

Each segment may in turn be split as well. This process yields a hierarchical rhetorical structure, i.e., an RST tree, for the analyzed piece of text. Each segment in this tree is either a nucleus or a satellite. Nuclei form the core of a text, whereas satellites support the nuclei and are considered to be less important for understanding a text. Several types of relations exist between RST elements. A satellite may, e.g., elaborate on or form a contrast with matters presented in a nucleus. A better understanding of a text's conveyed sentiment can be obtained by differentiating between text segments, based on such rhetorical roles (Heerschop et al., 2011a; Hogenboom et al., 2010b).

7.3 Classifying Polarity with Structure-Based Vector Representations of Text

As rule-based polarity classification has recently been shown to benefit from structure-guided sentiment analysis methods (Chenlo et al., 2013; Heerschop et al., 2011a; Hogenboom et al., 2015b; Taboada et al., 2008), we propose to harvest information from structural aspects of content in order to further improve the alternative, machine learning approach to polarity classification. To this end, we propose to classify the polarity of natural language text by using vector representations of text that incorporate not only word-based and sentiment-related features, but structure-based features as well. Linguistic processing of a document is required in order to be able to characterize it by means of such features.

7.3.1 Linguistic Processing

Our framework, visualized in Figure 7.1, takes several steps in order to enable the extraction of features that can be used by a machine learning classifier in order to classify the polarity of a document. First, we split a document into paragraphs and, subsequently, sentences and words. Then, for each sentence, we determine the Part-of-Speech (POS) and lemma of each word. Based on the identified POS and lemma, the word sense of each word is subsequently disambiguated by means of an algorithm that iteratively selects the word sense with the highest semantic similarity to the word's context (Heerschop et al., 2011a). In this word sense disambiguation process, we link the identified word senses to a semantic lexical resource, i.e., WordNet (Fellbaum, 1998). WordNet is organized into sets of cognitive synonyms – synsets – which can be differentiated based on their POS type. Each out of 117,659 synsets in WordNet expresses a distinct concept and may be linked to other synsets through various types of relations, e.g., synonymy or antonymy.

Having completed these preprocessing steps, we analyze the sentiment conveyed by the document's words, given their respective POS, lemma, and sense. To this end, we retrieve the sentiment score associated with each word's POS, lemma, and word sense from a sentiment lexicon, i.e., SentiWordNet 3.0 (Baccianella et al., 2010), which contains positivity, negativity, and objectivity scores for each synset in WordNet. We use this information to compute sentiment scores for each word by subtracting its associated negativity score from its associated positivity score, thus yielding a real number in the interval $[-1, 1]$, representing sentiment scores in the range from very negative to very positive, respectively.

In our analysis of the sentiment conveyed by the words constituting a document, we assign a weight to each word. These weights default to 1, but can be updated if the sentiment associated with specific words is detected to be negated or amplified. Following recent findings (Hogenboom et al., 2011a), we account for negation by inverting the polarity of the two words following a negation keyword that is listed in an existing negation lexicon (Hogenboom et al., 2011a), by multiplying their associated weights with -1. We account for amplification by means of an existing amplification lexicon, listing amplification keywords and their effect on the sentiment conveyed by the first succeeding word (Taboada et al., 2011).

One of the final steps in our framework for feature extraction for polarity classification involves identifying text segments and their respective rhetorical roles. In order to achieve this, we follow existing work (Chenlo et al., 2013; Heerschop et al., 2011a; Hogenboom et al., 2015b; Taboada et al., 2008) by segmenting the document's text in accordance with the top-level splits of sentence-level RST trees as generated by means of the SPADE parser (Soricut and Marcu, 2003). Furthermore, we allow for the most fine-grained analysis of the text by performing an additional segmentation in accordance with the leaf-level splits of the sentence-level RST trees generated by the SPADE parser.

The information thus obtained can subsequently be used in order to quantify the sentiment conveyed by (parts of) a document d. We define the sentiment score ζ_{s_i} of a segment s_i as the sum of the sentiment ζ_{t_j} associated with each word t_j in segment s_i, weighted with a weight w_{t_j} associated with these respective words, i.e.,

$$\zeta_{s_i} = \sum_{t_j \in s_i} \left(\zeta_{t_j} \cdot w_{t_j} \right), \quad \forall s_i \in R_d, \tag{7.1}$$

with R_d representing either all top-level or all leaf-level RST nodes in the sentence-level RST trees for document d, in case of top-level or leaf-level RST-guided sentiment analysis, respectively.

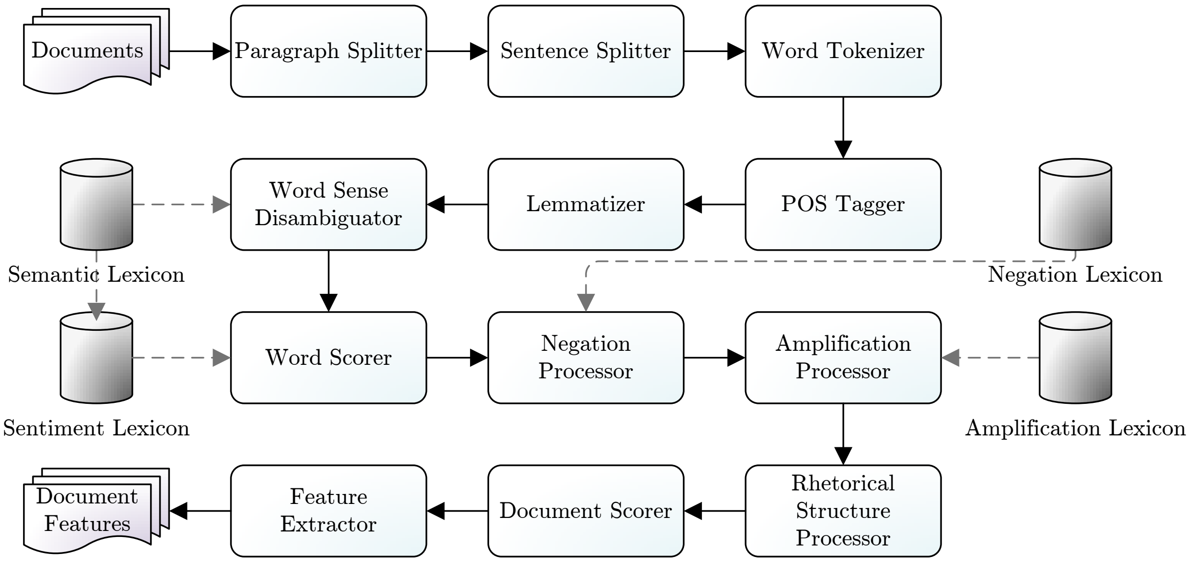


Figure 7.1: Overview of our sentiment analysis feature extraction framework. Solid arrows signal the information flow, whereas dashed arrows indicate a used-by relationship.

The segment-level scores thus computed can subsequently be aggregated in a document-level sentiment score ζ_d, i.e.,

$$\zeta_d = \sum_{s_i \in R_d} \zeta_{s_i}. \tag{7.2}$$

Once a document's sentiment score has been computed, the document has been fully processed. The results of the analysis can then be used for extracting features that characterize the document in a way that allows for the document's polarity to be determined.

7.3.2 Extracted Features

As discussed in Section 7.2.2, common, valuable features to be included in a vector representation of a document capture the (frequencies of) occurrence of specific words. These words could be simple lexical representations (i.e., strings of characters), or more complex ones, such as WordNet synsets. Inspired by the state-of-the-art (see Section 7.2.2), we use both representations. First, we represent text by means of the WordNet synsets (unigrams) that can be identified in its contents, as these synsets capture semantics and can be differentiated by their POS. Second, we represent text by means of its constituent lemmas (unigrams and bigrams), differentiated by their POS, in order to cover words that do not have an entry in WordNet. Another word-based feature extracted by our framework, based on the findings discussed in Section 7.2.2, is the length of a document, expressed in terms of its total number of words.

Our analysis of related work in Section 7.2.2 shows that other common, useful features relate to the sentiment conveyed by the text, as determined by means of a sentiment lexicon. Therefore, our framework extracts sentiment-related features that include the number of positive words, the number of negative words, and the sentiment scores of the sentiment-carrying words in a document, aggregated by means of (7.1) and (7.2). As the related research endeavors discussed in Section 7.2.2 have shown that information on negation and amplification is valuable when representing sentiment-carrying content in vector representations of text, we construct our sentiment-related features when performing four distinct types of sentiment analysis. We construct our sentiment-related word counts and scores when performing sentiment analysis without accounting for negation and amplification, sentiment analysis accounting for negation, sentiment analysis accounting for amplification, and sentiment analysis accounting for negation and amplification.

The sentiment-related features extracted by our framework can be used to characterize documents as a whole, but we propose to apply them to each distinct type of rhetorical element as well. Here, we define a rhetorical element as a text segment that has been identified as a nucleus or satellite belonging to a type of rhetorical relation on a specific level of analysis. A rhetorical element may for instance be an attributing satellite or the nucleus of a contrasting relation, in either the top-level split or a leaf-level split of a sentence-level RST tree. Our framework constructs features that capture the total number of words, the number of positive and negative words, and aggregated sentiment scores of the text segments that have been identified as specific rhetorical elements. The element-level features thus constructed allow for words and their conveyed sentiment to be treated differently, depending on their identified rhetorical role.

7.4 Evaluation

We evaluate the usefulness of our proposed (structure-based) features for polarity classification by means of a set of experiments. The setup of these experiments is detailed in Section 7.4.1. Additionally, we present our experimental results and discuss some caveats with respect to our findings in Sections 7.4.2, 7.4.3, and 7.4.4.

7.4.1 Experimental Setup

We evaluate the performance of various combinations of our considered features in a binary polarity classification task on two collections of documents. The first text collection consists of 1,000 positive and 1,000 negative English movie reviews (Pang and Lee, 2004).

The second corpus is a multi-domain collection of 8,000 English reviews, consisting of 1,000 positive and 1,000 negative reviews for each out of four distinct product categories, i.e., books, DVDs, electronics, and kitchen appliances (Blitzer et al., 2007).

Feature extraction for these corpora is performed by means of a Java-based implementation of our proposed framework for feature extraction. The initial tokenization steps in this implementation vary for our considered review corpora. For the movie review data, we detect paragraphs by making use of the `<P>` and `</P>` tags in the original HTML files of the reviews, as these tags signal the starts and the ends of paragraphs, respectively. In order to segment the identified paragraphs into sentences, we rely on the preprocessing done by Pang and Lee (2004). Conversely, for the multi-domain review corpus, we detect paragraphs by considering white lines to separate paragraphs. The paragraphs thus identified are split into sentences by means of the Stanford CoreNLP toolkit (Manning et al., 2014). For both corpora, we employ the Stanford Tokenizer (Manning et al., 2010) for identifying words in the identified sentences.

In order to identify the POS and lemma of each word thus identified, we use the OpenNLP (Baldridge and Morton, 2004) POS tagger and the Java WordNet Library (JWNL) API (Walenz and Didion, 2008), respectively. Only those words occurring in WordNet are actually lemmatized, whereas the lemma of each other word is in fact its original form. We link the words' senses to WordNet (Fellbaum, 1998) and retrieve their sentiment scores from SentiWordNet 3.0 (Baccianella et al., 2010). Furthermore, we account for negation by inverting the polarity of the two words following a negation keyword that is listed in a negation lexicon (Hogenboom et al., 2011a). We account for amplification by means of an amplification lexicon, listing amplification keywords and their effect on the sentiment conveyed by the first succeeding word (Taboada et al., 2011). Last, the rhetorical roles of words are identified by analyzing the top-level and leaf-level splits of sentence-level RST trees as generated by SPADE (Soricut and Marcu, 2003).

The implementation of our feature extraction framework allows us to compare the performance of machine learning models that use various sets of features in order to represent the documents in our corpora. These experiments are described in Section 7.4.1.1.

7.4.1.1 Experiments

We consider seven sets in three categories, i.e., four sets of word-based features, one set of sentiment-related features, and two sets of RST-based features (see Table 7.1). We assess the merits of each set individually, as well as in combination with other sets, with each combination containing at most one set from each category. Evaluating the performance of these combinations helps us assess the added value of each individual set of features.

Set	Type	Description
$\mathcal{B}$	Words	Document-level, binary features that indicate the presence of synset unigrams, inherently differentiated by POS.
$\mathcal{F}$	Words	Document-level indicators of the frequencies of occurrence of synset unigrams, inherently differentiated by POS.
$\mathcal{N}$	Words	Document-level, binary features indicating the presence of lemma n-grams, i.e., unigrams and bigrams that differentiate lemmas by POS.
$\mathcal{W}$	Words	Document-level indicators of the frequencies of occurrence of lemma n-grams, i.e., unigrams and bigrams that differentiate lemmas by POS.
$\mathcal{S}$	Sentiment	Document-level features capturing the number of words, the number of positive words, the number of negative words, and the sentiment scores for four types of sentiment analysis.
$\mathcal{T}$	RST	RST-based sentiment-related features, capturing the total, positive, and negative word counts, and the sentiment scores for four types of analysis, per top-level RST element type.
$\mathcal{L}$	RST	RST-based sentiment-related features, capturing the total, positive, and negative word counts, and the sentiment scores for four types of analysis, per leaf-level RST element type.

Table 7.1: The feature sets used in our experiments.

The word-based feature sets $\mathcal{B}$ and $\mathcal{F}$ contain features that indicate the respective presence and frequencies of occurrence of all WordNet synsets that occur in at least 5% of our data, i.e., 997 synsets for the movie review corpus, and 322 synsets for the multi-domain corpus. We apply this filter in order to keep the number of features tractable – considering all WordNet synsets would result in 117,659 features. Moreover, even though rare terms may be useful indicators for subjectivity (Wiebe et al., 2004), excluding such terms can yield models that generalize comparably well.

Similarly, the word-based feature sets $\mathcal{N}$ and $\mathcal{W}$ encompass features indicating the respective presence and frequencies of occurrence of all POS-specific lemma unigrams and bigrams that occur in at least 5% of our data, i.e., 1,157 n-grams for the movie review corpus, and 388 n-grams for the multi-domain corpus. This vastly reduces the feature space of 524,855 and 425,320 initially extracted n-grams for the movie review corpus and multi-domain review corpus, respectively.

Set $\mathcal{S}$ contains 16 features that capture the sentiment conveyed by the full text of our reviews. These features represent the sentiment score, the total word count, the number of positive words, and the number of negative words, as obtained by performing document-level sentiment analysis without accounting for negation and amplification (SA), sentiment analysis accounting for negation (SA^-), sentiment analysis accounting for amplification (SA^+), and sentiment analysis accounting for negation and amplification ($SA^\pm$).

Relation	Satellite description
ATTRIBUTION	Clause containing reporting verbs or cognitive predicates related to reported messages presented in the nucleus.
BACKGROUND	Information helping a reader to sufficiently comprehend matters presented in the nucleus.
CAUSE	An event leading to a result presented in the nucleus.
COMPARISON	Examination of matters along with matters presented in the nucleus.
CONDITION	Hypothetical, future, or otherwise unrealized situations, the realization of which influences the realization of nucleus matters.
CONTRAST	Situations juxtaposed to and compared with situations in the nucleus, which are considered as mostly similar, yet different in a few respects.
ELABORATION	Additional detail about matters presented in the nucleus.
ENABLEMENT	Information increasing a reader's potential ability of performing actions presented in the nucleus.
EVALUATION	Evaluative comments about matters presented in the nucleus.
EXPLANATION	Justifications or reasons for situations presented in the nucleus.
JOINT	No specific rhetorical relation holds with matters in the nucleus.
MANNER-MEANS	Explains how or by which means matters presented in the nucleus have been done.
SAME-UNIT	Text segment of which the subordinate nucleus belongs to the same rhetorical unit as the nucleus.
TEMPORAL	Events with an ordering in time with respect to events in the nucleus.

Table 7.2: Most common relations of satellites to their nuclei, identified by SPADE.

The RST-based feature sets $\mathcal{T}$ and $\mathcal{L}$ each contain 480 features representing the same 16 sentiment-related concepts for rhetorical elements in top-level ($\mathcal{T}$) or leaf-level ($\mathcal{L}$) splits of sentence-level RST trees. They encompass the nucleus and satellite elements for each out of 14 rhetorical relations that occur in at least 5% of our data (see Table 7.2), as well as a nucleus and a satellite element representing all other nuclei and satellites.

We assess the performance of each of our (combined) feature sets in terms of the precision, recall, and F_1-score for positive and negative documents separately, as well as the overall accuracy and macro-level F_1-score. Precision is the proportion of the positively (negatively) classified documents that are in fact positive (negative), whereas recall is the proportion of the actual positive (negative) documents that are also classified as such. The F_1-score is the harmonic mean of precision and recall. The macro-level F_1-score is the arithmetic mean of the F_1-scores of the positive and negative documents, weighted for their relative frequencies. Accuracy is the proportion of correctly classified documents. We assess the statistical significance of performance differences by means of a paired two-sample two-tailed t-test.

The performance is assessed under 10-fold cross-validation. For the movie review data, we use the same folds as in Chapter 5. For the multi-domain review corpus, we randomly split the data per domain into ten balanced folds, with 100 positive and 100 negative reviews each. For each (combined) set of features, our evaluation procedure is as follows. For each fold, we first perform a feature selection procedure on the fold's training data (see Section 7.4.1.2). A machine learning classifier that uses the selected features in order to classify the polarity of text is subsequently trained on the training data, and we evaluate its document polarity classification performance on the fold's test data (see Section 7.4.1.3). For each corpus, the resulting performance measures are subsequently aggregated over all folds in order to assess the overall performance of our feature sets.

7.4.1.2 Feature Selection

When performing feature selection on a training set, we first remove the features that show no variation over the training instances, as these features contain no information that can be used to distinguish between positive and negative polarity. Then, we rank the remaining features by the absolute value of their (Pearson) correlation with the document polarity and select those features with an absolute Pearson correlation coefficient of 0.1 or higher, in order to keep only those features that are at least somewhat relevant.

The absolute value of the Pearson correlation coefficient is a widely used ranking criterion, which is applicable to binary, continuous, and even (disjunctively coded) categorical features and target variables (Guyon, 2008). Our considered features are both binary and continuous, whereas our target variable, i.e., document polarity, is a categorical variable. As such, the absolute Pearson correlation coefficient is an attractive feature selection criterion for our data. The alternative wrapper methods for feature selection are less suitable in our particular case, due to the inherent computational complexity involved with evaluating the performance of the combinatorial explosion of subsets of features that can be constructed from our feature sets.

7.4.1.3 Polarity Classification

Using only those features selected by means of the procedure described in Section 7.4.1.2, we train a machine learning classifier on a training set and evaluate its polarity classification performance on a test set. In this work, we use an SVM classifier, as such classifiers are typically used in polarity classification tasks (Taboada et al., 2011). We use the WEKA (Hall et al., 2009) implementation of an SVM classifier, i.e., the *SMO* classifier, with a Radial Basis Function (RBF) kernel.

Two parameters of this classifier can be optimized, i.e., the parameters C and γ, both of which capture a trade-off between the complexity of the decision surface and the misclassification of training instances. A decision surface that is too complex may result in overfitting, so optimizing these parameters is of paramount importance. Therefore, before training our final classifier, we optimize the parameters of the classifier on the training data by means of a 10-fold cross-validated grid search procedure.

Our three-step parameter optimization procedure aims to find the values for C and γ that give the best accuracy on the training set, as assessed by means of internal 10-fold cross-validation. In the first step of our procedure, we perform a grid search on a logarithmic grid with base 10, with values of $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$ for both C and γ. Then, we perform a second grid search on a logarithmic grid with base 1.5, between the grid points surrounding the optimum found in the first iteration. Last, we perform a grid search between the grid points surrounding the optimum found in the second iteration, on a logarithmic grid with base 1.05.

After having optimized the C and γ parameters of our SVM classifier on the training set by means of our 10-fold cross-validated grid search procedure, we configure our classifier with the optimized parameters and train it on the full training set. Last, we evaluate the polarity classification performance of the trained classifier on the test set.

7.4.2 Experimental Results on Movie Reviews

The machine learning classifiers that use our various sets of features exhibit several trends in terms of polarity classification performance, as discussed in Section 7.4.2.1. The features selected by our machine learning polarity classifiers are analyzed in Section 7.4.2.2.

7.4.2.1 Polarity Classification Performance

The various combinations of features used in our machine learning models result in the polarity classification performance statistics reported in Table 7.3 and Figures 7.2 and 7.3. These results indicate that well-balanced and well-performing polarity classifiers can be trained when using (a combination of) our word-based, sentiment-related, and RST-based feature sets. Some of our polarity classifiers do however exhibit a marginally better performance on negative texts than they do on positive texts, which renders them somewhat less well-balanced. Additionally, the overall polarity classification performance of our classifiers shows a rather large variation over the utilized feature sets. The overall accuracy and macro-level F_1-scores on the movie review data range from about 65% for the worst-performing classifiers to about 82% for the best-performing ones.

	Positive			Negative			Overall	
Features	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy	F_1
$\mathcal{B}$	0.778	0.771	0.774	0.773	0.780	0.777	0.776	0.775
$\mathcal{F}$	0.777	0.754	0.765	0.761	0.784	0.772	0.769	0.769
$\mathcal{N}$	0.784	0.786	0.785	0.785	0.783	0.784	0.785	0.784
$\mathcal{W}$	0.814	0.800	0.807	0.803	0.817	0.810	0.809	0.808
$\mathcal{S}$	0.667	0.585	0.623	0.630	0.708	0.667	0.647	0.645
$\mathcal{T}$	0.674	0.640	0.657	0.657	0.691	0.674	0.666	0.665
$\mathcal{L}$	0.669	0.622	0.645	0.647	0.692	0.669	0.657	0.657
$\mathcal{BS}$	0.781	0.779	0.780	0.780	0.782	0.781	0.781	0.780
$\mathcal{BT}$	0.792	0.790	0.791	0.791	0.793	0.792	0.792	0.791
$\mathcal{BL}$	0.781	0.783	0.782	0.783	0.781	0.782	0.782	0.782
$\mathcal{FS}$	0.790	0.773	0.781	0.778	0.794	0.786	0.784	0.783
$\mathcal{FT}$	0.796	0.775	0.785	0.781	0.801	0.791	0.788	0.788
$\mathcal{FL}$	0.789	0.772	0.780	0.777	0.793	0.785	0.783	0.782
$\mathcal{NS}$	0.799	0.813	0.806	0.810	0.796	0.803	0.805	0.804
$\mathcal{NT}$	0.807	0.798	0.802	0.800	0.809	0.805	0.804	0.803
$\mathcal{NL}$	0.798	0.816	0.807	0.812	0.793	0.802	0.805	0.804
$\mathcal{WS}$	0.818	0.810	0.814	0.812	0.820	0.816	0.815	0.815
$\mathcal{WT}$	0.823	0.814	**0.819**	**0.816**	0.825	**0.820**	**0.820**	**0.819**
$\mathcal{WL}$	0.825	0.809	0.817	0.813	0.828	0.820	0.819	0.818
$\mathcal{ST}$	0.686	0.633	0.658	0.659	0.710	0.684	0.672	0.671
$\mathcal{SL}$	0.681	0.646	0.663	0.663	0.698	0.680	0.672	0.672
$\mathcal{BST}$	0.791	0.791	0.791	0.791	0.791	0.791	0.791	0.791
$\mathcal{BSL}$	0.781	0.779	0.780	0.780	0.782	0.781	0.781	0.780
$\mathcal{FST}$	0.793	0.772	0.782	0.778	0.798	0.788	0.785	0.785
$\mathcal{FSL}$	0.787	0.774	0.781	0.778	0.791	0.784	0.783	0.782
$\mathcal{NST}$	0.809	0.801	0.805	0.803	0.811	0.807	0.806	0.806
$\mathcal{NSL}$	0.807	**0.817**	0.812	0.815	0.805	0.810	0.811	0.811
$\mathcal{WST}$	0.823	0.807	0.815	0.811	0.827	0.819	0.817	0.817
$\mathcal{WSL}$	**0.826**	0.803	0.814	0.808	**0.831**	0.820	0.817	0.817

Table 7.3: The 10-fold cross-validated performance of our feature sets on the movie review corpus. The best performance is printed in bold for each performance measure.

The worst-performing classifiers use only the sentiment-based features in $\mathcal{S}$, the leaf-level RST-based features in $\mathcal{L}$, the top-level RST-based features in $\mathcal{T}$, or a combination of these sets, i.e., $\mathcal{SL}$ or $\mathcal{ST}$. However, the features in $\mathcal{S}$, $\mathcal{L}$, and $\mathcal{T}$ become particularly useful once combined with the comparably well-performing word-based features in $\mathcal{B}$, $\mathcal{F}$, $\mathcal{N}$, and especially $\mathcal{W}$. Our best classifiers include RST-based features, sometimes combined with document-level sentiment-related features from $\mathcal{S}$. For instance, our three best-performing classifiers – which do not significantly differ from one another in terms of overall classification performance – use feature set combinations $\mathcal{WT}$, $\mathcal{WL}$, and $\mathcal{WST}$.

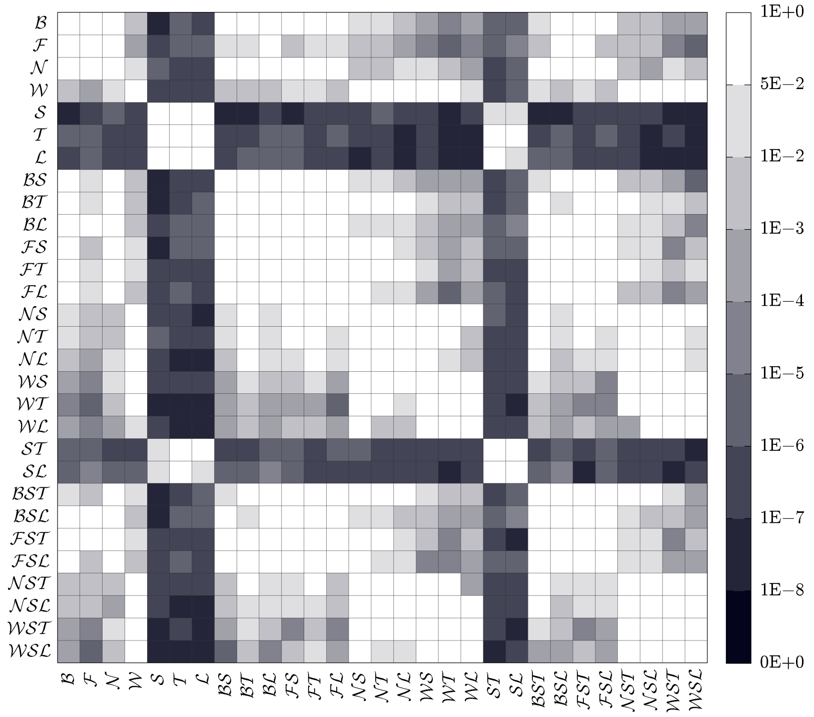


Figure 7.2: The p-values for the paired, two-tailed t-test assessing the statistical significance of differences in mean accuracy obtained by using our (combined) feature sets on the movie review corpus.

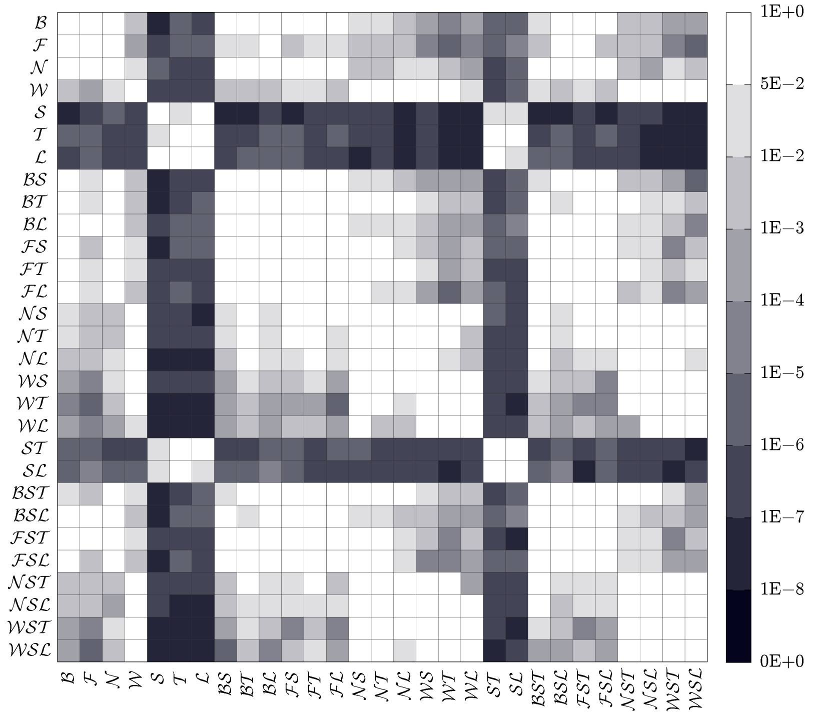


Figure 7.3: The p-values for the paired, two-tailed t-test assessing the statistical significance of differences in mean macro-level F_1-scores obtained by using our (combined) feature sets on the movie review corpus.

	Accuracy				F_1			
Features	$+\mathcal{B}$	$+\mathcal{F}$	$+\mathcal{N}$	$+\mathcal{W}$	$+\mathcal{B}$	$+\mathcal{F}$	$+\mathcal{N}$	$+\mathcal{W}$
$\mathcal{S}$	0.207	0.212	0.244	0.261	0.210	0.214	0.247	0.263
$\mathcal{T}$	0.189	0.184	0.207	0.231	0.190	0.184	0.208	0.232
$\mathcal{L}$	0.190	0.191	0.225	0.246	0.191	0.192	0.225	0.247
$\mathcal{ST}$	0.178	0.169	0.200	0.217	0.179	0.170	0.201	0.218
$\mathcal{SL}$	0.161	0.164	0.207	0.216	0.162	0.165	0.207	0.216

Table 7.4: Relative change in the 10-fold cross-validated overall performance on the movie review corpus when including word-based features (significant at $p < 0.0001$).

Other observations that can be made from Table 7.3 and Figures 7.2 and 7.3 relate to performance differences between similar feature sets. Our results show that feature sets that include binary synset features from $\mathcal{B}$ or frequency-based lemma features from $\mathcal{W}$ tend to perform better than their respective counterparts that include frequency-based synset features from $\mathcal{F}$ or binary lemma features from $\mathcal{N}$. However, these performance differences are mostly statistically insignificant. Similarly, top-level RST-based features in $\mathcal{T}$ appear to be associated with a better overall polarity classification performance than leaf-level RST-based features in $\mathcal{L}$, but these performance differences are not statistically significant either. On the other hand, lemma-based features from sets $\mathcal{N}$ and $\mathcal{W}$ tend to yield significantly better polarity classification performance than synset-based features from sets $\mathcal{B}$ and $\mathcal{F}$. Because the general purpose WordNet synsets do not cover all words occurring in the movie reviews, our lemma-based features can represent the movie reviews' content more accurately, thus facilitating a more accurate polarity classification.

In general, individual feature sets, i.e., $\mathcal{B}$, $\mathcal{F}$, $\mathcal{N}$, $\mathcal{W}$, $\mathcal{S}$, $\mathcal{T}$, and $\mathcal{L}$, tend to perform better once they are combined with one another – the classifiers that use features from multiple feature sets exhibit the best 10-fold cross-validated performance in our experiments. Tables 7.4, 7.5, and 7.6 provide insight into the effects of combining word-based, sentiment-related, and RST-based features, respectively, with one another.

Table 7.4 clearly shows that adding word-based features from sets $\mathcal{B}$, $\mathcal{F}$, $\mathcal{N}$, or $\mathcal{W}$ to sentiment-related or RST-based features yields vast, significant performance improvements of up to about 26% in terms of overall accuracy and macro-level F_1-scores. The performance improvements drop to about 20% when adding word-based features to combined sets of sentiment-related and RST-based features, as these richer representations of natural language text already allow for a better distinction between positive and negative documents than the $\mathcal{S}$, $\mathcal{L}$, and $\mathcal{T}$ sets individually (see Table 7.3). The observed added value of word-based features confirms their substantial importance for polarity classification purposes, as suggested in related work discussed in Section 7.2.2.

	Accuracy	F_1
Features	$+\mathcal{S}$	$+\mathcal{S}$
$\mathcal{B}$	0.006	0.006
$\mathcal{F}$	0.019^{**}	0.019^{**}
$\mathcal{N}$	0.025^{**}	0.025^{**}
$\mathcal{W}$	0.008	0.008
$\mathcal{T}$	0.009	0.009
$\mathcal{L}$	0.023^{*}	0.023^{*}
$\mathcal{BT}$	-0.001	-0.001
$\mathcal{BL}$	-0.002	-0.002
$\mathcal{FT}$	-0.004	-0.004
$\mathcal{FL}$	0.000	0.000
$\mathcal{NT}$	0.003	0.003
$\mathcal{NL}$	0.008	0.008
$\mathcal{WT}$	-0.003	-0.003
$\mathcal{WL}$	-0.002	-0.002

Table 7.5: Relative change in the 10-fold cross-validated overall performance on the movie review corpus when including sentiment-related features. Performance differences marked with * are statistically significant at $p < 0.05$, and those marked with ** are statistically significant at $p < 0.01$.

	Accuracy		F_1	
Features	$+\mathcal{T}$	$+\mathcal{L}$	$+\mathcal{T}$	$+\mathcal{L}$
$\mathcal{B}$	0.021	0.008	0.021	0.008
$\mathcal{F}$	0.025^{*}	0.018^{*}	0.025^{*}	0.018^{*}
$\mathcal{N}$	0.024^{**}	0.025^{*}	0.024^{**}	0.025^{*}
$\mathcal{W}$	0.014	0.012^{*}	0.014	0.012^{*}
$\mathcal{S}$	0.039^{*}	0.039^{*}	0.040^{*}	0.041^{*}
$\mathcal{BS}$	0.013^{*}	0.000	0.013^{*}	0.000
$\mathcal{FS}$	0.002	-0.001	0.002	-0.001
$\mathcal{NS}$	0.002	0.008	0.002	0.008
$\mathcal{WS}$	0.002	0.002	0.002	0.002

Table 7.6: Relative change in the 10-fold cross-validated overall performance on the movie review corpus when including RST-based features. Performance differences marked with * are statistically significant at $p < 0.05$, and those marked with ** are statistically significant at $p < 0.01$.

The added value of the sentiment-related features in $\mathcal{S}$ is more limited, as exhibited by Table 7.5. Adding sentiment-related features to word-based or RST-based features can yield modest overall performance improvements of up to about 2%, which are statistically significant for frequency-based synset features $\mathcal{F}$, binary lemma-based n-grams $\mathcal{N}$, and leaf-level RST-based features $\mathcal{L}$. Adding sentiment-related features from $\mathcal{S}$ to combined word-based and RST-based features does not yield any significant performance improvements. This suggests that the document-level sentiment-related information in feature set $\mathcal{S}$ does not add much to the information that is already covered by the well-performing combinations of word-based features with our novel RST-based features that capture sentiment-related information on the level of rhetorical elements.

Adding RST-based features to word-based and document-level sentiment-related features yields mostly significant, yet modest improvements in overall performance of up to approximately 4%, as indicated by Table 7.6. The RST-based sentiment-related information in feature sets $\mathcal{T}$ and $\mathcal{L}$ has the most convincing added value over the document-level sentiment-related information captured by feature set $\mathcal{S}$. Yet, the features in $\mathcal{T}$ and – to a lesser extent – $\mathcal{L}$ have some added value over word-based features as well. For instance, adding RST-based information to frequency-based synset features from feature set $\mathcal{F}$ or to binary lemma-based n-grams from feature set $\mathcal{N}$ yields small yet significant overall polarity classification performance improvements between 2% and 3%. Furthermore, the 2% improvements in accuracy and macro-level F_1-scores obtained by adding feature set $\mathcal{T}$ to the binary synset-based feature set $\mathcal{B}$ are just short of qualifying as statistically significant, with respective p-values of 0.052 and 0.050. Nevertheless, the introduction of top-level RST-based features $\mathcal{T}$ to a combination of binary synset-based features $\mathcal{B}$ and document-level sentiment-related features $\mathcal{S}$ does in fact yield a small, significant improvement in overall polarity classification performance of about 1%.

All in all, the inclusion of word-based features in our machine learning polarity classifier seems to have the most impact on the overall polarity classification performance on our considered corpus of movie reviews. However, adding sentiment-related information, especially on the level of rhetorical elements, can yield modest, yet significant performance improvements as well – models that include such information generally significantly outperform their counterparts that do not include such information.

Interestingly, the document-level sentiment-related information captured by the features that constitute feature set $\mathcal{S}$ does not have much added value over RST-based sentiment-related information, and especially not when features from feature set $\mathcal{T}$ are used. Conversely, our RST-based features in set $\mathcal{L}$ and especially those in feature set $\mathcal{T}$ have a larger, significant added value over the features that constitute feature set $\mathcal{S}$.

With their sentiment-related information at the level of rhetorical elements, feature sets $\mathcal{L}$ and $\mathcal{T}$ can capture much of the document-level sentiment-related information that set $\mathcal{S}$ captures, whereas the features in set $\mathcal{S}$ cannot express all information captured by feature sets $\mathcal{L}$ and $\mathcal{T}$. These findings render our novel RST-based features – especially those capturing sentiment-related information for the top-level splits of sentence-level RST trees – the most fruitful additions to word-based features.

7.4.2.2 Selected Features

The polarity classification performance reported in Section 7.4.2.1 is not based on all extracted features that constitute the feature sets used by our classifiers, but rather on comparably small subsets of features, that have been selected by means of the feature selection procedure described in Section 7.4.1.2. The feature counts in Table 7.7 show that on average, only about 8% of all extracted features is selected. The only exception here is our smallest feature set, i.e., $\mathcal{S}$, where 75% of all extracted features is selected.

Our comparably well-performing classifiers generally use more features (in absolute terms) than the classifiers that exhibit a less competitive performance. Nevertheless, using more features does not guarantee a better performance. Our best-performing classifiers use on average 137, 132, and 149 features from the $\mathcal{WT}$, $\mathcal{WL}$, and $\mathcal{WST}$ sets, respectively, whereas some other classifiers perform worse while using a similar or even higher number of features. Clearly, the quality of features is important as well. The features used in our best-performing models can shed more light onto what information is truly useful in automated polarity classification. These models' most important features – i.e., those most strongly correlated with document polarity – exhibit several patterns, as demonstrated by Figures 7.4, 7.5, and 7.6.

The characteristics of the single most important feature selected for each out of ten folds for our three best-performing sets of features, i.e., $\mathcal{WT}$, $\mathcal{WL}$, and $\mathcal{WST}$, are visualized in Figure 7.4. In 33% of the cases, the single most important feature selected by our classifiers is a feature capturing information related to document-level sentiment, whereas in 67% of the cases, the most important feature captures sentiment-related information on the level of rhetorical elements. Interestingly, word presence or frequencies do not turn out to be among the single most important features, in spite of their strong and significant impact on the performance of our classifiers, as discussed in Section 7.4.2.1. An explanation for this phenomenon lies in the comparably complex nature of our sentiment-related features, which condense a lot of information related to how specific words are used in order to convey sentiment.

	Extracted	Selected	
Features	Total	μ	σ
$\mathcal{B}$	997	58	4.206
$\mathcal{F}$	997	65	4.261
$\mathcal{N}$	1,157	76	3.225
$\mathcal{W}$	1,157	88	4.104
$\mathcal{S}$	16	12	0.000
$\mathcal{T}$	480	48	3.072
$\mathcal{L}$	480	43	4.627
$\mathcal{BS}$	1,013	70	4.206
$\mathcal{BT}$	1,477	107	6.652
$\mathcal{BL}$	1,477	101	7.632
$\mathcal{FS}$	1,013	77	4.261
$\mathcal{FT}$	1,477	113	5.980
$\mathcal{FL}$	1,477	108	7.245
$\mathcal{NS}$	1,173	88	3.225
$\mathcal{NT}$	1,637	124	5.678
$\mathcal{NL}$	1,637	119	5.934
$\mathcal{WS}$	1,173	100	4.104
$\mathcal{WT}$	1,637	137	6.258
$\mathcal{WL}$	1,637	132	6.341
$\mathcal{ST}$	496	60	3.072
$\mathcal{SL}$	496	55	4.627
$\mathcal{BST}$	1,493	119	6.652
$\mathcal{BSL}$	1,493	113	7.632
$\mathcal{FST}$	1,493	125	5.980
$\mathcal{FSL}$	1,493	120	7.245
$\mathcal{NST}$	1,653	136	5.678
$\mathcal{NSL}$	1,653	131	5.934
$\mathcal{WST}$	1,653	149	6.258
$\mathcal{WSL}$	1,653	144	6.341

Table 7.7: Feature counts for our feature sets on the movie review corpus, reported as their total number of extracted features, as well as the mean (μ) and standard deviation (σ) of the number of features selected in each individual fold.

Our models' most valuable document-level sentiment-related features capture the reviews' lexicon-based sentiment scores. These scores stem from the method discussed in Section 7.3.1 and account for both negation and amplification ($SA^{\pm}$). The most important RST-based features capture similar sentiment scores, computed for some nuclei of the top-level and leaf-level splits of sentence-level RST trees. These nuclei do not belong to the 14 most salient rhetorical relations, but capture the combined nuclei of all other rhetorical relations, and thus cover the core information for many rhetorical roles at once.

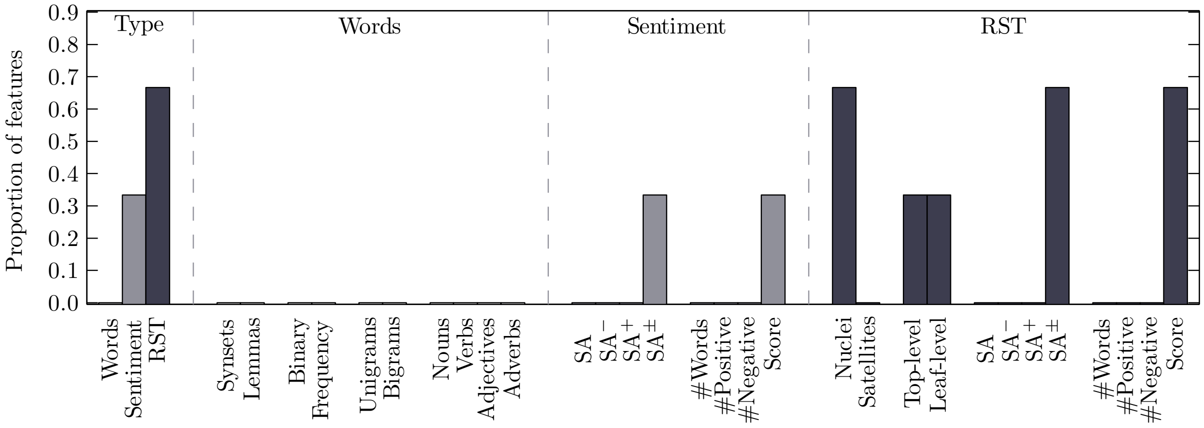


Figure 7.4: Characteristics of the top 1 features selected for all folds of our three best-performing feature sets on the movie review corpus, i.e., $\mathcal{WT}$, $\mathcal{WL}$, and $\mathcal{WST}$.

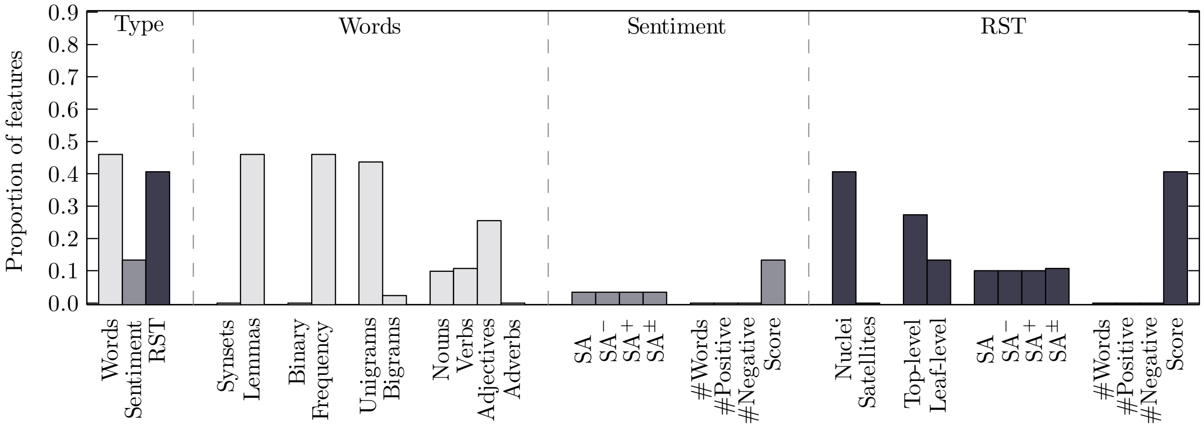


Figure 7.5: Characteristics of the top 10 features selected for all folds of our three best-performing feature sets on the movie review corpus, i.e., $\mathcal{WT}$, $\mathcal{WL}$, and $\mathcal{WST}$.

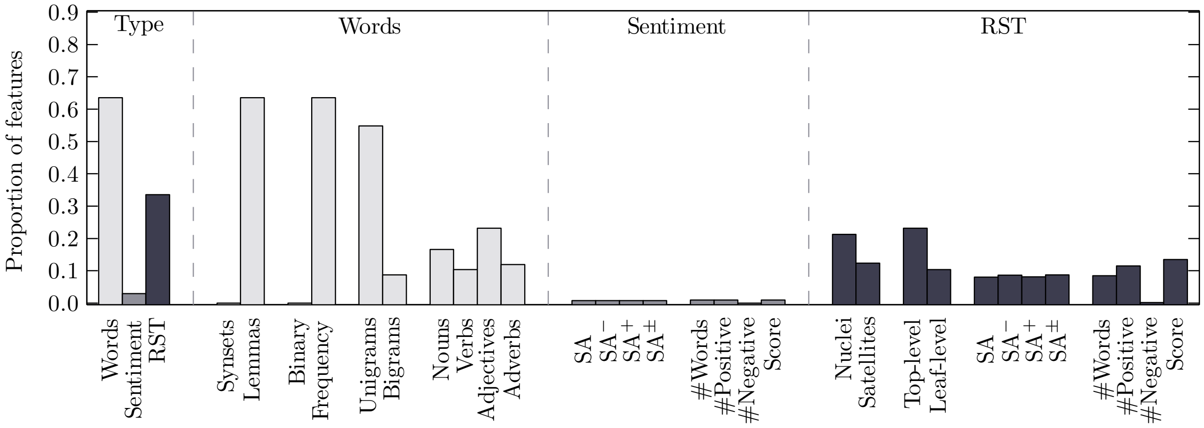


Figure 7.6: Characteristics of all features selected for all folds of our three best-performing feature sets on the movie review corpus, i.e., $\mathcal{WT}$, $\mathcal{WL}$, and $\mathcal{WST}$.

Figure 7.5 demonstrates the more varied nature of the ten best features that have been selected for each fold for our best-performing feature sets. Of these features, 46% is word-based, 13% is sentiment-related, and another 41% is RST-based. This suggests that specific words used in our reviews are important indicators for the polarity of these reviews, while a small majority of the most important features captures sentiment-related information, mostly on the level of RST elements.

Word-based features included in the ten best features of our models using the $\mathcal{WT}$, $\mathcal{WL}$, and $\mathcal{WST}$ feature sets are frequencies of lemmas that represent words that mostly express negative opinions. The most useful lemmas are typically adjectives, such as "*bad*" (also in combination with the noun "*movie*"), "*ridiculous*", "*stupid*", and "*great*". The nouns "*mess*" and "*life*" are valuable indicators for the polarity of a review as well. The high discriminative power of the word "*life*" may be domain- or even corpus-specific, as in our corpus, quite a few positive reviews describe how well a movie captures (the struggles, challenges, or absurdity of) real life. Another rather peculiar, yet important feature turns out to be the presence of the verb "*to suppose*". In our corpus, reviewers often use this verb in order to express that they are not too convinced about something, or – more often than not – that their expectations have not been met. Examples of these uses can be found in the phrases "*Her side-kick was supposed to be funny but just annoyed me*" and "*The film is supposed to be a big suspense thriller, but (...) the film never achieves suspense, or even a sense of intrigue*". Another verb that can act as a proxy for a review's polarity is the verb "*to waste*", which is typically used in order to express a perceived waste of money or talent.

The document-level sentiment-related features in the top ten features of our best-performing models cover sentiment scores computed by performing sentiment analysis without accounting for negation or amplification (SA), or by performing a type of sentiment analysis that accounts for negation (SA^-), amplification (SA^+), or both negation and amplification ($SA^\pm$). A similar pattern can be observed for the RST-based sentiment-related features in the top ten features of our models. These features relate to (mostly top-level) nuclei only and cover – besides the nuclei covered by the single best features – the nuclei of JOINT relations, which occur in almost every review and as such cover the sentiment conveyed by an alleged core part of many reviews.

Figure 7.6 shows that even in all features selected by the models based on our three best-performing feature sets, sentiment-related information is valuable, especially in case this information is RST-based. Word-based features cover 63% of all selected features, whereas document-level sentiment-related features and RST-based sentiment-related features cover 3% and 34% of the features, respectively.

Besides the words covered by the top ten features, the word-based features selected by our best-performing models cover the frequencies of occurrence of the lemmas of many adjectives, adverbs, nouns, and verbs. The numerous additional adjectives include "*awful*", "*terrible*, "*boring*", "*predictable*", "*memorable*", "*effective*", "*complex*", "*intelligent*", "*hilarious*", "*good*" (combined with the noun "*film*"), "*perfect*", and "*excellent*". Additional adverbs include "*unfortunately*", "*perfectly*", and "*well*". Noteworthy additional nouns for the movie review data include "*nothing*" (e.g., "*Nothing in this movie makes sense*"), "*flaw*", "*plot*" (typically used when addressing flaws in the plot), and "*performance*" (typically used in order to express that an actor delivered quite a performance). The sentiment-carrying verbs "*to deserve*" and "*to fail*" are used by our best-performing models as well.

All sentiment-related features used in our best-performing models cover sentiment scores, total word counts, and positive word counts as obtained by performing our four considered sentiment analysis variants, i.e., those that do and those that do not account for negation and/or amplification. The RST-based sentiment-related features do however exhibit a slight tendency of favoring sentiment analysis variants that at least account for negation over variants that do not take into account negation. The RST-based sentiment-related features cover rhetorical relations in mostly top-level splits of sentence-level RST trees. Most of these features cover nuclei, but some satellites are represented too. This suggests that satellites – which are considered to contain less relevant information – in fact contain useful information that can help distinguish positive from negative texts.

Satellites that elaborate on information presented in nuclei, i.e., ELABORATION satellites, turn out to be important features. The persuasive ENABLEMENT satellites are important too. Additionally, our best-performing models often include features that capture the sentiment in ATTRIBUTION satellites. These satellites present the context of messages reported in nuclei, and are apparently more important than the reported message itself. Consider, for example, the phrase "*Any studio executive that thinks this plot is going to win points with the reviewing press needs to check into rehab*". In this phrase, the reported message of the plot being praised by the reviewing press is subordinate to its negative context, where the reporting verb (i.e., "*to think*" or "*to suppose*") in itself already has some negative connotations in this context. Another important satellite turns out to be the CONDITION satellite, which provides crucial contextual information for matters presented in nuclei. An example of this phenomenon can be found in the phrase "*We wouldn't mind a minute of Johnny Mnemonic if the action played better*", where the nucleus suggests a positive sentiment with respect to the movie, whereas the satellite clarifies that this would only hold if it were not for the lousy action.

Overall, in our best-performing polarity classification models, sentiment-carrying words – especially adjectives – turn out to be valuable features. Our best classifiers mostly use features that capture the frequency of occurrence of specific lemmas (predominantly unigrams). The most valuable information, however, appears to be derived from sentiment-related, and mostly RST-based features. Especially nuclei of top-level splits of sentence-level RST trees turn out to contain valuable cues for the polarity of movie reviews, yet some types of satellites that provide crucial contextual information play an important role as well. These observations suggest that features that capture sentiment information, especially when related to the structure of documents, form a valuable addition to commonly used word-based features.

7.4.3 Experimental Results on Multi-Domain Reviews

Our experimental results on the multi-domain review corpus of Blitzer et al. (2007) largely confirm the experimental results on the movie review corpus of Pang and Lee (2004), presented in Section 7.4.2. Our main findings on the multi-domain review collection are discussed in detail in Sections 7.4.3.1 and 7.4.3.2.

7.4.3.1 Polarity Classification Performance

The various combinations of features used in our machine learning models result in the overall, cross-domain polarity classification performance statistics reported in Table 7.8 and Figures 7.7 and 7.8. As is the case for the movie review corpus, well-balanced and well-performing polarity classifiers can be trained for the multi-domain corpus by using (a combination of) our word-based, sentiment-related, and RST-based feature sets. However, our classifiers for the multi-domain review corpus show a smaller variation over the utilized feature sets in terms of the overall polarity classification performance than our movie review classifiers do. The overall accuracy and macro-level F_1-scores on the multi-domain review data range from approximately 70% to 78%.

The worst-performing classifiers use (combinations of) the sentiment-based features in $\mathcal{S}$, the leaf-level RST-based features in $\mathcal{L}$, and the top-level RST-based features in $\mathcal{T}$. However, these features become particularly useful once combined with the comparably well-performing word-based features in $\mathcal{B}$, $\mathcal{F}$, $\mathcal{N}$, and especially $\mathcal{W}$. Combinations of feature sets typically yield a better overall polarity classification performance than each feature set individually. Our best classifiers include top-level RST-based features, sometimes combined with document-level sentiment-related features. For instance, our three best-performing classifiers use feature set combinations $\mathcal{WST}$, $\mathcal{WT}$, and $\mathcal{WSL}$.

	Positive			Negative			Overall	
Features	Precision	Recall	F_1	Precision	Recall	F_1	Accuracy	F_1
$\mathcal{B}$	0.714	0.715	0.714	0.715	0.714	0.715	0.715	0.714
$\mathcal{F}$	0.740	0.712	0.726	0.700	0.729	0.715	0.720	0.720
$\mathcal{N}$	0.742	0.730	0.736	0.726	0.737	0.732	0.734	0.734
$\mathcal{W}$	0.759	0.737	0.748	0.729	0.752	0.740	0.744	0.744
$\mathcal{S}$	0.712	0.699	0.705	0.693	0.706	0.700	0.702	0.702
$\mathcal{T}$	0.709	0.708	0.709	0.708	0.709	0.708	0.709	0.708
$\mathcal{L}$	0.712	0.705	0.708	0.703	0.709	0.706	0.707	0.707
$\mathcal{BS}$	0.759	0.755	0.757	0.754	0.758	0.756	0.756	0.756
$\mathcal{BT}$	0.770	0.765	0.767	0.763	0.768	0.766	0.766	0.766
$\mathcal{BL}$	0.763	0.753	0.758	0.750	0.759	0.755	0.756	0.756
$\mathcal{FS}$	0.760	0.758	0.759	0.757	0.759	0.758	0.758	0.758
$\mathcal{FT}$	0.760	0.765	0.762	0.766	0.761	0.764	0.763	0.763
$\mathcal{FL}$	0.757	0.766	0.761	0.768	0.760	0.764	0.763	0.763
$\mathcal{NS}$	0.772	0.758	0.765	0.753	0.767	0.760	0.763	0.762
$\mathcal{NT}$	0.774	0.770	0.772	0.768	0.773	0.771	0.771	0.771
$\mathcal{NL}$	0.773	0.769	0.771	0.768	0.771	0.770	0.770	0.770
$\mathcal{WS}$	**0.792**	0.769	0.780	0.762	**0.785**	0.773	0.777	0.777
$\mathcal{WT}$	0.787	0.775	0.781	0.772	0.784	0.778	0.779	0.779
$\mathcal{WL}$	0.781	0.773	0.777	0.771	0.778	0.774	0.776	0.776
$\mathcal{ST}$	0.714	0.713	0.713	0.712	0.713	0.713	0.713	0.713
$\mathcal{SL}$	0.710	0.705	0.708	0.703	0.708	0.705	0.707	0.706
$\mathcal{BST}$	0.768	0.765	0.766	0.764	0.767	0.765	0.766	0.766
$\mathcal{BSL}$	0.765	0.763	0.764	0.762	0.764	0.763	0.763	0.763
$\mathcal{FST}$	0.767	0.770	0.768	0.771	0.768	0.769	0.769	0.769
$\mathcal{FSL}$	0.762	0.765	0.763	0.767	0.763	0.765	0.764	0.764
$\mathcal{NST}$	0.777	0.774	0.775	0.773	0.776	0.774	0.775	0.775
$\mathcal{NSL}$	0.772	0.768	0.770	0.767	0.770	0.769	0.769	0.769
$\mathcal{WST}$	0.786	**0.779**	**0.783**	**0.777**	0.784	**0.780**	**0.782**	**0.781**
$\mathcal{WSL}$	0.785	0.772	0.779	0.769	0.782	0.775	0.777	0.777

Table 7.8: The 10-fold cross-validated performance of our feature sets on the multi-domain corpus. The best performance is printed in bold for each performance measure.

Our experimental results show that frequency-based word features from sets $\mathcal{F}$ and $\mathcal{W}$ tend to yield a better overall polarity classification performance than binary word-based features from sets $\mathcal{B}$ and $\mathcal{N}$. Moreover, the lemma-based variants $\mathcal{N}$ and $\mathcal{W}$ tend to outperform the synset-based variants $\mathcal{B}$ and $\mathcal{F}$. Additionally, top-level RST-based features $\mathcal{T}$ appear to be associated with a better overall polarity classification performance than leaf-level RST-based features $\mathcal{L}$. However, these observed performance differences between similar word-based or RST-based feature sets are more often than not statistically insignificant for the multi-domain review corpus.

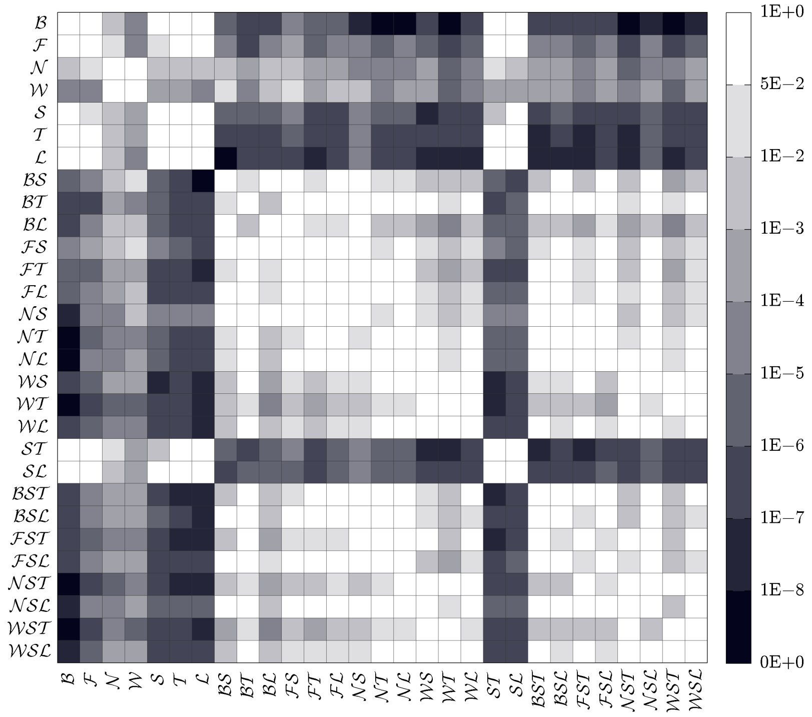


Figure 7.7: The p-values for the paired, two-tailed t-test assessing the statistical significance of differences in mean accuracy obtained by using our (combined) feature sets on the multi-domain review corpus.

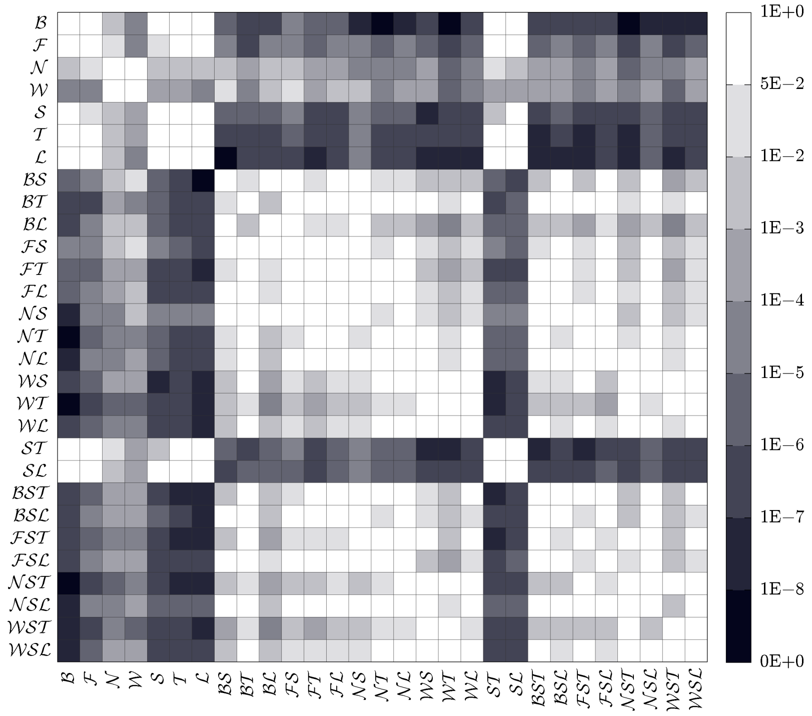


Figure 7.8: The p-values for the paired, two-tailed t-test assessing the statistical significance of differences in mean macro-level F_1-scores obtained by using our (combined) feature sets on the multi-domain review corpus.

	Accuracy				F_1			
Features	$+\mathcal{B}$	$+\mathcal{F}$	$+\mathcal{N}$	$+\mathcal{W}$	$+\mathcal{B}$	$+\mathcal{F}$	$+\mathcal{N}$	$+\mathcal{W}$
$\mathcal{S}$	0.077	0.080	0.086	0.106	0.077	0.080	0.086	0.106
$\mathcal{T}$	0.082	0.077	0.089	0.100	0.082	0.077	0.089	0.100
$\mathcal{L}$	0.069	0.078	0.089	0.097	0.069	0.078	0.089	0.097
$\mathcal{ST}$	0.074	0.078	0.087	0.096	0.074	0.078	0.087	0.096
$\mathcal{SL}$	0.081	0.081	0.089	0.100	0.081	0.081	0.089	0.100

Table 7.9: Relative change in the 10-fold cross-validated overall performance on the multi-domain corpus when including word-based features (significant at $p < 0.0001$).

Combining the predictive power of various types of features can have a significant, positive impact on the overall polarity classification performance, as demonstrated by Tables 7.9, 7.10, and 7.11. Clearly, the most substantial, significant performance improvements can be obtained by adding synset-based features or especially lemma-based features to sets of sentiment-related features. Doing so can yield overall polarity classification performance improvements between about 7% and 10% on the multi-domain review corpus. This suggests that the words used in order to convey opinions contain important information. Yet, it is also important to capture how these words convey sentiment. This is demonstrated by the substantial, significant performance improvements that can be obtained when enriching word-based information with document-level and especially RST-based sentiment-related information. On the multi-domain review collection, these improvements can amount to almost 6% for document-level sentiment-related information, and over 7% for RST-based features. It should however be noted that document-level sentiment-based information does not have a significant added value over RST-based features, and that RST-based features in turn appear to have only limited added value over document-level sentiment-based information. This renders RST-based sentiment-related information the most fruitful addition to word-based features.

7.4.3.2 Selected Features

Table 7.12 demonstrates that the observed polarity classification performance on the multi-domain review corpus realized by means of our considered feature sets is in fact obtained by using only a small selection of features from these sets. Typically, only about 5% of the extracted features is actually used in our classifiers. Furthermore, larger sets of selected features (in absolute terms) generally tend to result in models that exhibit a comparably good polarity classification performance. The characteristics of the features selected from the feature sets that yield the best performance, i.e., $\mathcal{WST}$, $\mathcal{WT}$, and $\mathcal{WSL}$, are visualized in Figures 7.9, 7.10, and 7.11.

	Accuracy	F_1
Features	$+\mathcal{S}$	$+\mathcal{S}$
$\mathcal{B}$	0.058***	0.058***
$\mathcal{F}$	0.053***	0.053***
$\mathcal{N}$	0.039***	0.039***
$\mathcal{W}$	0.044***	0.044***
$\mathcal{T}$	0.006	0.006
$\mathcal{L}$	-0.001	-0.001
$\mathcal{BT}$	-0.001	-0.001
$\mathcal{BL}$	0.009**	0.009**
$\mathcal{FT}$	0.008*	0.008*
$\mathcal{FL}$	0.002	0.002
$\mathcal{NT}$	0.005*	0.005*
$\mathcal{NL}$	-0.001	-0.001
$\mathcal{WT}$	0.003	0.003
$\mathcal{WL}$	0.002	0.002

Table 7.10: Relative change in the 10-fold cross-validated overall performance measures on the multi-domain corpus when including sentiment-related features. Performance differences marked with * are statistically significant at $p < 0.05$, those marked with ** are significant at $p < 0.01$, and those marked with *** are significant at $p < 0.001$.

	Accuracy		F_1	
Features	$+\mathcal{T}$	$+\mathcal{L}$	$+\mathcal{T}$	$+\mathcal{L}$
$\mathcal{B}$	0.073***	0.058***	0.073***	0.058***
$\mathcal{F}$	0.059***	0.059***	0.059***	0.059***
$\mathcal{N}$	0.051***	0.050***	0.051***	0.050***
$\mathcal{W}$	0.048***	0.043***	0.048***	0.043***
$\mathcal{S}$	0.015**	0.006	0.015**	0.006
$\mathcal{BS}$	0.012**	0.009	0.012**	0.009
$\mathcal{FS}$	0.014*	0.007	0.014*	0.007
$\mathcal{NS}$	0.016**	0.009	0.016**	0.009
$\mathcal{WS}$	0.006	0.000	0.006	0.000

Table 7.11: Relative change in the 10-fold cross-validated overall performance measures on the multi-domain corpus when including RST-based features. Performance differences marked with * are statistically significant at $p < 0.05$, those marked with ** are significant at $p < 0.01$, and those marked with *** are significant at $p < 0.001$.

	Extracted	Selected	
Features	Total	μ	σ
$\mathcal{B}$	322	18	2.000
$\mathcal{F}$	322	17	2.654
$\mathcal{N}$	388	27	4.389
$\mathcal{W}$	388	26	4.295
$\mathcal{S}$	16	7	1.786
$\mathcal{T}$	480	13	2.242
$\mathcal{L}$	480	14	3.452
$\mathcal{BS}$	338	24	2.479
$\mathcal{BT}$	802	31	2.745
$\mathcal{BL}$	802	31	2.846
$\mathcal{FS}$	338	24	3.401
$\mathcal{FT}$	802	31	3.190
$\mathcal{FL}$	802	31	3.275
$\mathcal{NS}$	404	34	5.752
$\mathcal{NT}$	841	40	4.684
$\mathcal{NL}$	841	41	6.022
$\mathcal{WS}$	377	33	5.739
$\mathcal{WT}$	868	39	4.722
$\mathcal{WL}$	868	40	6.239
$\mathcal{ST}$	496	20	3.116
$\mathcal{SL}$	496	21	4.821
$\mathcal{BST}$	818	38	3.343
$\mathcal{BSL}$	818	38	4.287
$\mathcal{FST}$	818	38	4.025
$\mathcal{FSL}$	818	38	4.836
$\mathcal{NST}$	884	47	6.105
$\mathcal{NSL}$	884	48	7.630
$\mathcal{WST}$	884	46	6.187
$\mathcal{WSL}$	884	47	7.845

Table 7.12: Feature counts for our feature sets on the multi-domain review corpus, reported as their total number of extracted features, as well as the mean (μ) and standard deviation (σ) of the number of features selected in each individual fold.

In 78% of the cases, the single most important features selected by our best classifiers capture sentiment scores on the level of rhetorical elements (see Figure 7.9). The remaining 22% captures document-level sentiment scores. As is the case for the movie review corpus, the presence of specific words is not among the most important proxies for the polarity of a review in the multi-domain review corpus. Again, an explanation for this can be found in the richness of our sentiment-related features, as these features capture how sentiment is conveyed by *all* words occurring in (a rhetorical element of) a review.

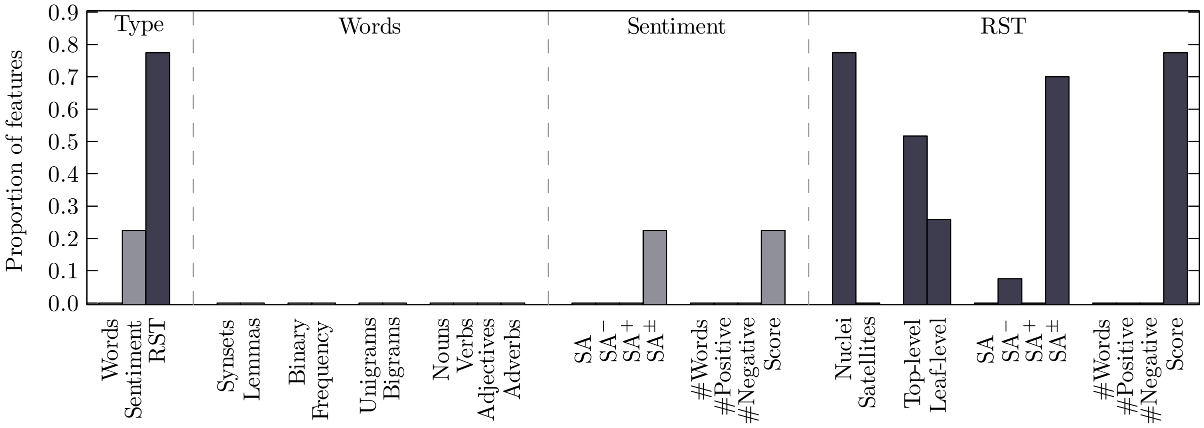


Figure 7.9: Characteristics of the top 1 features selected for all folds of our three best-performing feature sets on the multi-domain review corpus, i.e., $\mathcal{WST}$, $\mathcal{WT}$, and $\mathcal{WSL}$.

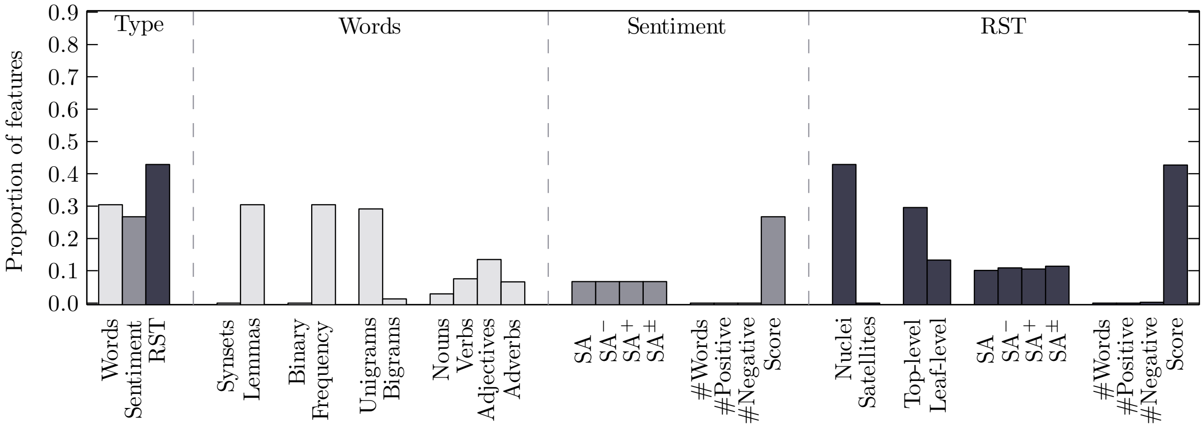


Figure 7.10: Characteristics of the top 10 features selected for all folds of our three best-performing feature sets on the multi-domain review corpus, i.e., $\mathcal{WST}$, $\mathcal{WT}$, and $\mathcal{WSL}$.

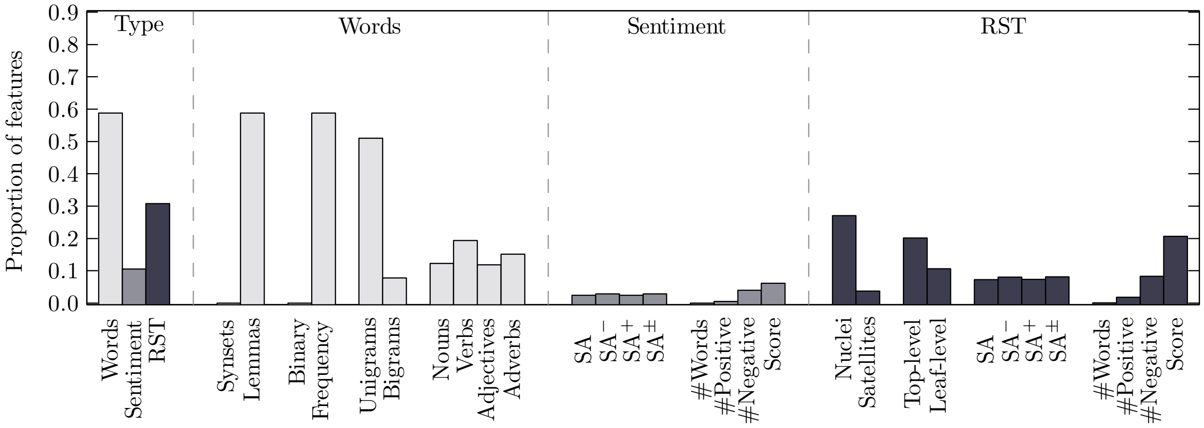


Figure 7.11: Characteristics of all features selected for all folds of our three best-performing feature sets on the multi-domain review corpus, i.e., $\mathcal{WST}$, $\mathcal{WT}$, and $\mathcal{WSL}$.

The sentiment-related information captured by the single most informative features selected by each of our best-performing classifiers always accounts for negation, and often takes into account amplification of sentiment as well. Those features that capture sentiment-related information for distinct rhetorical elements do so exclusively for the combined nuclei of rhetorical relations that are not among the most salient ones in our data. These particular features cover a large variety of types of core information at once, and as such contain a lot of – apparently comparably useful – information.

Sentiment-related information, especially the RST-based variant, dominates the top ten features selected by our best polarity classifiers as well, as demonstrated by Figure 7.10. Document-level sentiment-related features cover 27% of the top ten selected features and RST-based sentiment-related information is represented by another 43% of the top ten selected features, whereas the remaining 30% consists of word-based features.

The word-based features that are among the top ten features selected by our polarity classifiers that use the $\mathcal{WST}$, $\mathcal{WT}$, and $\mathcal{WSL}$ feature sets represent lemmas of mostly positive words. Most of these words are adjectives, representing the usual suspects "*easy*", "*good*", "*great*", "*excellent*", "*perfect*", and "*bad*". An interesting informative word turns out to be the adverb "*not*", sometimes preceded by the verb "*to do*". This word in itself does not carry any sentiment, but it rather negates the sentiment conveyed by other words. In the multi-domain corpus, "*(do) not*" occurs notably more often in negative reviews than in positive reviews – negative opinions in this corpus often tend to be expressed or even emphasized by negating the opposite. Other informative verbs turn out to be "*to enjoy*", "*to love*", and "*to return*", the latter of which is often used in a negative context, e.g., in order to express that the reviewed item was or should be returned to the store. Noteworthy selected nouns include "*money*" and "*price*". In the multi-domain review corpus, "*money*" is much more likely to be used in a negative context, e.g., in order to express that a product is a waste of money. Conversely, "*price*" is much more likely to be used in a positive context, e.g., in order to express that a product is attractively priced.

The document-level and RST-based sentiment-related features in the top ten features selected from the $\mathcal{WST}$, $\mathcal{WT}$, and $\mathcal{WSL}$ feature sets cover sentiment scores computed by performing sentiment analysis without accounting for negation or amplification, or by performing a type of sentiment analysis that accounts for negation, amplification, or both negation and amplification. Furthermore, RST-based sentiment-related features among the ten most informative features of each of our three best-performing models represent exclusively nuclei. These are (mostly top-level) nuclei that stem from Joint relations, in addition to the nuclei covered by the single most useful features. Joint relations occur in many reviews and as such cover a substantial part of the core content of many reviews.

Figure 7.11 shows that even in all features selected by our best-performing models, sentiment-related information is valuable, especially when this information is RST-based. Nevertheless, word-based features cover a small majority of all selected features, i.e., 59%. Document-level sentiment-related features and RST-based sentiment-related features cover 10% and 31% of the features, respectively.

Besides the words covered by the top ten selected features, the word-based features selected by our best classifiers cover the frequencies of occurrence and – to a lesser extent – binary indicators of the presence of the lemmas of many adjectives, adverbs, nouns, and verbs. The additional adjectives include "*nice*" and "*little*". The latter adjective is typically used in terms of endearment (e.g., "*I love this little thing*" or "*This little gem*"), or in order to downplay negative aspects of a product in an otherwise positive review (e.g., "*The soup bowls are a little on the small side*"). Notable additional adverbs include "*well*" (typically used in a positive context) and "*instead*" (in order to, e.g., express a mismatch between expectations and reality). Other words selected by our best classifiers include the nouns "*love*", "*service*", and "*support*", the latter two of which are especially valuable proxies for negative sentiment in the electronics domain, where needing support turns out to be a good indicator for bad product experiences. Last, noteworthy additional verbs include "*to recommend*" and "*to be*", combined with numerous positive and negative adjectives like "*great*" and "*bad*".

All sentiment-related features used in our best-performing models cover sentiment scores and negative word counts as obtained by performing any of our considered sentiment analysis variants, but preferably by means of a variant that at least accounts for negation of the sentiment conveyed by specific words. The RST-based sentiment-related features cover rhetorical relations in mostly top-level splits of sentence-level RST trees. Most of these features cover nuclei, with Elaboration, Condition, and Attribution satellites forming the exception. This suggests that, on the multi-domain review corpus, the sentiment-related information in satellites is of limited use for polarity classification, whereas the sentiment conveyed by nuclei tends to be rather useful when classifying reviews in various domains as either positive or negative.

All in all, as is the case for the movie review corpus, sentiment-carrying words turn out to be indispensable to good polarity classifiers for the multi-domain review corpus. Yet, sentiment-related information, especially when guided by RST, turns out to contain valuable additional cues for a review's polarity. Using this type of information in a machine learning polarity classifier in addition to traditional word-based features enables significant improvements in polarity classification performance.

7.4.4 Caveats

Our experimental results presented in Sections 7.4.2 and 7.4.3 show that the addition of structure-based features to traditional word-based features is invariably useful across collections of reviews in various domains. However, even though our evaluation yields promising and encouraging results, several caveats should be taken into consideration.

First, some of our considered word-based features are linked to the semantic categories in a general purpose semantic lexicon, i.e., to synsets in WordNet. As explained in Section 7.3.2, such a representation enables us to capture the semantics and POS information of words, thus allowing for more robust models. However, the WordNet synsets may not cover all lexical representations of words occurring in a corpus. Highly domain-specific words may not be covered either. This explains why the word-based features that are based on lexical representations of (the lemmas of) words tend to yield a better polarity classification performance. The trade-off between robustness and domain-specificity may affect the quality of the document-level and RST-based sentiment-related features as well, as these features rely on the SentiWordNet 3.0 sentiment lexicon, which only contains sentiment scores for each synset in WordNet.

Another caveat is related to our feature selection process. We disregard features that occur in only a small part of our corpora, even though these features could be valuable (Wiebe et al., 2004). Moreover, we disregard features that are hardly correlated with the polarity class of the reviews in our corpora. This methodology can be justified as it allows us to reduce the dimensionality of our data and to make our models less prone to overfitting. However, other subsets of features may exist that yield an even better polarity classification performance than the performance reported in Sections 7.4.2 and 7.4.3. These subsets may be found by using other feature selection methods, for instance by means of genetic algorithms or ant colony optimization techniques that evaluate many different feature subsets in order to identify the best subset. However, the computational complexity of training our non-linear classifiers forms a major bottleneck here, thus rendering such wrapper methods unfeasible in our current setup.

7.5 Conclusions

In this chapter, we have demonstrated how machine learning approaches to polarity classification of text can benefit from novel features that capture structural aspects of natural language text. Typical machine learning approaches heavily rely on the presence of specific (groups of) words and as such inherently focus on *what* is said in a piece of text.

However, as recent advances in rule-based sentiment analysis suggest that it may be more important *how* sentiment-carrying words are used in a text (as signaled by the text's rhetorical structure), we have proposed features that capture the sentiment of distinct rhetorical elements in a text, and we have evaluated the usefulness of these features in machine learning polarity classifiers for collections of English reviews in various domains.

Our experimental results over 10,000 English reviews suggest that the *what* and the *how* are both important cues for the polarity of natural language text. Word-based features are indispensable to good polarity classifiers, yet (mostly) structure-based sentiment information provides valuable additional guidance that can help significantly improve the polarity classification performance of machine learning classifiers. In fact, the most informative features used by our best-performing classifiers capture the sentiment conveyed by specific rhetorical elements. Most of these elements constitute the core of a text, yet some elements provide crucial contextual information that is not typically considered to be part of a text's core, but rather of its supporting content that is generally deemed predominantly irrelevant for conveying sentiment.

Thus, we have successfully applied recent findings for rule-based sentiment analysis to a performance-wise more competitive machine learning approach to sentiment analysis. Our proposed richer vector representation of natural language text contributes to more effective automated sentiment analysis systems that can help better support decision making processes that require accurate insight into one's stakeholders' sentiment. Our findings, however, warrant several directions for future research.

A first direction for future research could be to validate our findings in other domains. Second, other feature selection mechanisms and classifiers could be explored in order to further improve the performance of our current models. Third, future work could focus on exploring how the full rhetorical context of words – denoted by the full paths of rhetorical relations from the root node of a rhetorical structure tree to leaf nodes that represent text segments – can be captured in a vector representation of text. Last, the *what* and the *how* could be combined in future work, by differentiating word presence by rhetorical elements. For our current corpora, this is infeasible because of the data sparsity issues that arise due to the high dimensionality of our data, compared to the number of instances in our corpora. The usefulness of such features would hence need to be evaluated on a larger corpus. Such experiments would require classifiers and feature selection mechanisms that can handle the inherently substantially larger amount of data – with a much higher dimensionality – in a computationally efficient and effective way.

Chapter 8

Conclusions and Outlook

The work presented in this dissertation explores how the automated analysis of the sentiment conveyed by natural language text can be performed more effectively by utilizing higher levels of linguistic analysis than existing methods do. Whereas typical approaches to automated sentiment analysis are confined to determining the polarity of text by making use of mostly morphological, lexical, and syntactic information, this dissertation shows that additional analyses of semantics and discourse structure can yield substantially more effective systems. Such systems exploit more of the potential of information contained within natural language text, by accounting for the semantic context and rhetorical roles of cues for sentiment that are identified by means of morphological, lexical, and syntactic analysis. The main findings that constitute this conclusion are summarized in Section 8.1. Sections 8.2 and 8.3 provide an outlook on the implications of the encouraging findings presented in this dissertation, as well as on possibly fruitful directions for future research.

8.1 Main Findings

The first finding brought forward by the research efforts underlying this dissertation is that the morphological, lexical, and syntactic information traditionally used in automated sentiment analysis can help distinguish between polarity classes only to a limited extent. The constructed sentiment mappings that support this finding model the relation between sentiment scores based on low-level linguistic analysis on the one hand, and universal classes of authors' intended sentiment on the other hand. Interestingly, the nature of these mappings appears to differ across collections of documents written in various languages. This suggests that the way in which people express their sentiment may be context-dependent, thus rendering polarity classification solely based on low-level linguistic analysis particularly challenging.

Fortunately, some people use emoticons as additional cues that signal the sentiment that they intend to convey through their language. Empirical results reported in this dissertation indicate that people typically use emoticons in natural language text in order to express, stress, or disambiguate their sentiment in particular text segments, thus rendering emoticons helpful local proxies for the sentiment conveyed by a text as a whole. Because of the interactions that emoticons have on a semantic level with other words in the text segments in which they occur, the sentiment associated with these emoticons tends to dominate the sentiment conveyed by the words, mostly on a paragraph level. Modeling the mechanics of these semantic interactions in a rule-based polarity classifier yields a significantly improved polarity classification performance, compared to using low-level linguistic analysis only.

Accounting for semantics when analyzing the polarity of text turns out to have a significant added value in a multi-lingual polarity classification setting as well. An intuitive method would involve machine-translating text to a reference language and subsequently analyzing the polarity of the translated text by means of a polarity classifier that has been developed for the reference language. This approach turns out to be rather unsuccessful, as machine translation may yield an inaccurate representation of the original content in the reference language, and semantics may be lost in translation. A significantly better polarity classification performance can be achieved when exploiting semantic relations between and within languages in order to construct a sentiment lexicon for a new language, based on an existing sentiment lexicon for the reference language, or on a sufficiently large and diverse set of seed words for the new language.

Another important finding is that a better understanding of a text's conveyed sentiment can be obtained by performing linguistic analysis not only on a morphological, lexical, syntactic, and semantic level, but also on a discourse level. Automated sentiment analysis should be guided by a deep and fine-grained analysis of a text's rhetorical structure, as characterized by (automatically) applying the Rhetorical Structure Theory (RST) of Mann and Thompson (1988). This enables a rule-based polarity classifier to make a fine-grained distinction between important text segments and less important ones in terms of their contribution to a text's overall sentiment, based on their rhetorical roles. Such an approach can yield significant polarity classification performance improvements, compared to baselines not or only shallowly accounting for rhetorical structure. In an RST-guided polarity classifier, accounting for the rhetorical structure of individual sentences appears to yield better results than accounting for paragraph-level or document-level rhetorical structure, as the impact of an automated RST parser's occasional misclassifications of rhetorical relations is minimized for smaller units of analysis, such as sentences.

A drawback of RST-guided sentiment analysis is the computational complexity of the natural language processing techniques that automatically identify the rhetorical structure of a piece of text. However, the scalability of RST-guided sentiment analysis can be improved by performing a more focused analysis. This can be accomplished by guiding the sentiment analysis process by a shallow analysis of the sentential rhetorical structure of only a small fraction of all text. The experimental results presented in this dissertation suggest that an improved understanding of a carefully selected fraction of a text can already yield a significantly better understanding of the text as a whole, in terms of its conveyed sentiment.

Even though rule-based sentiment analysis methods allow for intuitive ways of incorporating discourse analysis into the process, machine learning approaches can benefit from a more sophisticated linguistic analysis as well. Features that stem from morphological, lexical, syntactic, or semantic linguistic analysis – mostly representing the occurrence of specific (senses of) words – are indispensable to good polarity classifiers. Yet, structure-based sentiment information can provide valuable additional guidance that can yield significant improvements in terms of polarity classification performance of a machine learning classifier. The most informative features identified in this dissertation capture the sentiment conveyed by specific rhetorical elements that either constitute a text's core or provide crucial contextual information.

All in all, the findings of the work covered by this dissertation suggest that the polarity of natural language text should not be determined solely based on *what* is said, but based on *how* the message of a text is conveyed as well. This can be achieved by accounting for the interactions of cues of sentiment on a semantic level, as well as by guiding the polarity analysis process by a text's rhetorical structure.

8.2 Future Work

This dissertation shows how a more effective analysis of the polarity of natural language text can be achieved by involving not only low-level linguistic processing, but also semantic analysis and discourse analysis. The incorporation of such high-level linguistic analyses facilitates the extraction of as much information as possible from the actual natural language content of a text. Yet, an even better understanding of a text may be achieved by considering the highest level of linguistic analysis, which deals with pragmatics and requires real-world knowledge in order to identify meaning that is not encoded in the text itself. The incorporation of this type of information into the analysis of the polarity of natural language text seems a viable, yet challenging direction for future work.

Semantic Web (Berners-Lee et al., 2001) technologies can be utilized in order to incorporate real-world knowledge into the analysis. Such technologies can be used to represent the semantic context of the concepts discussed in a text by linking these concepts to information in large publicly available semantic repositories that can be as general-purpose as DBpedia (Lehmann et al., 2015), or as domain-specific as the Linked Movie Data Base (Consens et al., 2008). Inference can subsequently be applied on the linked data in order to improve the understanding of the meaning of the considered text. Determining what real-world knowledge to include in which part of the analysis, as well as how and where to find this knowledge may prove to be far from trivial tasks. A particular challenge here lies in gathering the relevant linked data in the first place, as it needs to be effectively and efficiently retrieved from a multitude of vast, heterogeneous data sources (Hogenboom et al., 2009b,c, 2010c, 2012d, 2013b, 2015e).

Another issue that requires further investigation is the scalability of sentiment analysis approaches that are guided by high-level linguistic analyses of semantics, discourse, and possibly pragmatics. The findings of this dissertation suggest that especially the automated identification of rhetorical structure can be a computationally intensive process. However, promising results can be obtained with a more scalable approach that focuses structure-guided sentiment analysis on only a small set of carefully selected text segments that form good proxies for the subjective content of a text as a whole. Scalability issues could be further addressed by investigating the usefulness of other notions of structure, for instance based on discourse markers (Taboada, 2006), textual entailment (Herrera et al., 2006), or a small-world representation of text (Balinksy et al., 2011).

8.3 Outlook

The promising findings of this dissertation can be applied to other sentiment analysis tasks besides the polarity classification of full pieces of text. The identification of on-topic, polarized text segments (Mei et al., 2007; O'Hare et al., 2009; Zhang and Ye, 2008) may for instance benefit from an analysis of a text's rhetorical structure, as this structure gives clues on the role that each text segment plays in conveying the overall message of the text. Moreover, a more detailed analysis of a text's natural language content by accounting for semantics and discourse information may also enable a more fine-grained analysis of the sentiment conveyed by a text, with respect to, for instance, various aspects of an entity of interest (Boiy and Moens, 2009; Jiang et al., 2011; Schouten and Frasincar, 2014). This may prove helpful in opinion summarization tasks (Hu and Liu, 2004; Lerman et al., 2009; Mangnoesing et al., 2012; Titov and McDonald, 2008) as well.

By now, automated sentiment analysis techniques – especially approaches that include the higher levels of linguistic analysis dealt with in this dissertation – can provide for a good level of understanding of the sentiment conveyed by natural language content. The ever-growing computational power that is at our disposal facilitates the applicability of these techniques on an increasingly large scale. Thus, given the pivotal role of stakeholders' sentiment in today's business and economic processes, the high-level linguistics-based view on sentiment advocated by this dissertation should find its way to real-life decision support systems in order to better facilitate well-informed decision making in practice.

Promising applications lie in specific domains, in which a timely and accurate quantification of one's stakeholders' sentiment is warranted in order to enable effective support for decision making processes. Reputation management (Amigo et al., 2013; Jansen et al., 2009) is one of such typical application scenarios in which classical surveys may not fully address today's information needs in a timely and effective manner. Other potentially fruitful application scenarios for the findings presented in this dissertation include dynamic pricing in highly competitive and volatile markets with ever-changing circumstances (Hogenboom et al., 2009a, 2010a, 2015d), news-based algorithmic trading in stock markets (Nuij et al., 2014; Hogenboom et al., 2012e,f), and the construction of a more deliberate conceptualization of economic sentiment that complements traditional macro-economic indicators with a quantification of a general mood (Vuchelen, 2004). In each of these scenarios, automated sentiment analysis tools should take into account that sentiment may not so much be revealed by people's words per se, but rather by the way in which people use these words in order to convey their sentiment.

Bibliography

A. Abbasi, H. Chan, and A. Salem. Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. *ACM Transactions on Information Systems*, 26(3):1–34, 2008.

E. Amigo, J. Carrillo de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, T. Martin-Wanton, E. Meij, M. de Rijke, and D. Spina. Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. In *4th International Conference of the CLEF Initiative (CLEF 2013)*, volume 8138 of *Lecture Notes in Computer Science*, pages 333–352. Springer, 2013.

I. Arnold and E. Vrugt. Fundamental Uncertainty and Stock Market Volatility. *Applied Financial Economics*, 18(17):1425–1440, 2008.

H. Baazaoui Zghal, M. Aufaure, and N. Ben Mustapha. A Model-Driven Approach of Ontological Components for On-line Semantic Web Information Retrieval. *Journal of Web Engineering*, 6(4):309–336, 2007.

S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *7th Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2200–2204. European Language Resources Association, 2010.

N. Bach, M. Nguyen, and A. Shimazu. EDU-Based Similarity for Paraphrase Identification. In *18th International Conference on Applications of Natural Language to Information Systems (NLDB 2013)*, volume 7934 of *Lecture Notes in Computer Science*, pages 65–76. Springer, 2013.

D. Bal, M. Bal, A. van Bunningen, A. Hogenboom, F. Hogenboom, and F. Frasincar. Sentiment Analysis with a Multilingual Pipeline. In *12th International Conference on Web Information System Engineering (WISE 2011)*, volume 6997 of *Lecture Notes in Computer Science*, pages 129–142. Springer, 2011.

A. Balahur, J. Hermida, and A. Montoyo. Detecting Implicit Expressions of Emotion in Text: A Comparative Analysis. *Decision Support Systems*, 53(4):742–753, 2012.

J. Baldridge and T. Morton. OpenNLP, 2004. Available online, `http://opennlp.sourceforge.net/`.

H. Balinksy, A. Balinsky, and S. Simske. Document Sentences as a Small World. In *2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2011)*, pages 2583–2588. IEEE, 2011.

C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan. Multilingual Subjectivity Analysis Using Machine Translation. In *2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 127–135. Association for Computational Linguistics, 2008.

C. Banea, R. Mihalcea, and J. Wiebe. Multilingual Subjectivity: Are More Languages Better? In *23rd International Conference on Computational Linguistics (COLING 2010)*, pages 28–36. Association for Computational Linguistics, 2010.

D. Baron. Competing for the Public through the News Media. *Journal of Economics and Management Strategy*, 14(2):339–376, 2005.

M. Bautin, L. Vijayarenu, and S. Skiena. International Sentiment Analysis for News and Blogs. In *2nd International Conference on Weblogs and Social Media (ICWSM 2008)*, pages 19–26. Association for the Advancement of Artificial Intelligence, 2008.

T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284 (5):34–43, 2001.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 440–447. Association for Computational Linguistics, 2007.

E. Boiy and M. Moens. A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts. *Information Retrieval*, 12(5):526–558, 2009.

M. Bovi. Economic versus Psychological Forecasting. Evidence from Consumer Confidence Surveys. *Journal of Economic Psychology*, 30(4):563–574, 2009.

J. Burgoon, D. Buller, and W. Woodall. *Nonverbal Communication: The Unspoken Dialogue.* McGraw-Hill, 2nd edition, 1996.

E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.

L. Carlson, D. Marcu, and M. Okoruwski. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current Directions in Discourse and Dialogue*, pages 85–112. Kluwer Academic Publishers, 2003.

C. Cesarano, B. Dorr, A. Picariello, D. Reforgiato, A. Sagoff, and V. Subrahmanian. OASYS: An Opinion Analysis System. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs (CAAW 2006)*, pages 21–26. Association for the Advancement of Artificial Intelligence, 2006.

S. Chan. Beyond Keyword and Cue-Phrase Matching: A Sentence-Based Abstraction Technique for Information Extraction. *Decision Support Systems*, 42(2):759–777, 2006.

C. Chang, C. Hsu, and S. Lui. Automatic Information Extraction from Semi-Structured Web Pages by Pattern Discovery. *Decision Support Systems*, 35(1):129–147, 2003.

C. Chang, M. Kayed, R. Girgis, and K. Shaalan. A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1411–1428, 2006.

P. Chaovalit and L. Zhou. Movie Review Mining: A Comparison Between Supervised and Unsupervised Classification Approaches. In *38th Hawaii International Conference on System Sciences (HICSS 2005)*, page 112c. IEEE, 2005.

B. Chardon, F. Benamara, Y. Mathieu, V. Popescu, and N. Asher. Measuring the Effect of Discourse Structure on Sentiment Analysis. In *14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)*, volume 7817 of *Lecture Notes in Computer Science*, pages 25–37. Springer, 2013.

J. Chenlo and D. Losada. Effective and Efficient Polarity Estimation in Blogs Based on Sentence-Level Evidence. In *20th ACM Conference on Information and Knowledge Management (CIKM 2011)*, pages 365–374. Association for Computing Machinery, 2011.

J. Chenlo, A. Hogenboom, and D. Losada. Sentiment-Based Ranking of Blog Posts using Rhetorical Structure Theory. In *18th International Conference on Applications of Natural Language to Information Systems (NLDB 2013)*, volume 7934 of *Lecture Notes in Computer Science*, pages 13–24. Springer, 2013.

J. Chenlo, A. Hogenboom, and D. Losada. Rhetorical Structure Theory for Polarity Estimation: An Experimental Study. *Data and Knowledge Engineering*, 94:135–147, 2014.

T. Childers and M. Houston. Conditions for a Picture-Superiority Effect on Consumer Memory. *Journal of Consumer Research*, 11(2):643–654, 1984.

ComputerUser. Emoticons, 2013. Available online, `http://www.computeruser.com/resources/dictionary/emoticons.html`.

M. Consens, O. Hassanzadeh, and A. Teisanu. Linked Movie Data Base, 2008. Available online, `http://www.linkedmdb.org/`.

W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. TextFlow: Towards Better Understanding of Evolving Topics in Text. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2412–2421, 2011.

W. Dai, G. Xue, Q. Yang, and Y. Yu. Co-Clustering Based Classification for Out-of-Domain Documents. In *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, pages 210–219. Association for Computing Machinery, 2007a.

W. Dai, G. Xue, Q. Yang, and Y. Yu. Transferring Naive Bayes Classifiers for Text Classification. In *22nd AAAI Conference on Artificial Intelligence (AAAI 2007)*, pages 540–545. Association for the Advancement of Artificial Intelligence, 2007b.

T. Dao and T. Simpson. Measuring Similarity between Sentences. Technical report, WordNet.Net, 2005. Available online, `http://wordnetdotnet.googlecode.com/svn/trunk/Projects/Thanh/Paper/WordNetDotNet_Semantic_Similarity.pdf`.

D. Davidov, O. Tsur, and A. Rappoport. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In *23rd International Conference on Computational Linguistics (COLING 2010)*, pages 241–249. Association for Computational Linguistics, 2010.

N. Della Penna and H. Huang. Constructing Consumer Sentiment Index for U.S. Using Google Searches. Technical Report 2009-26, University of Alberta, 2009. Available online, `http://econpapers.repec.org/repec:ris:albaec:2009_026`.

A. Devitt and K. Ahmad. Sentiment Polarity Identification in Financial News: A Cohesion-based Approach. In *45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 984–991. Association for Computational Linguistics, 2007.

X. Ding, B. Lu, and P. Yu. A Holistic Lexicon-Based Approach to Opinion Mining. In *1st ACM International Conference on Web Search and Web Data Mining (WSDM 2008)*, pages 231–240. Association for Computing Machinery, 2008.

A. Esuli and F. Sebastiani. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *5th Conference on Language Resources and Evaluation (LREC 2006)*, pages 417–422. European Language Resources Association, 2006.

R. Feldman. Techniques and Applications for Sentiment Analysis. *Communications of the ACM*, 56(4):82–89, 2013.

C. Fellbaum. *WordNet: An Electronic Lexical Database.* MIT Press, 1998.

S. Gerani, M. Carman, and F. Crestani. Proximity-Based Opinion Retrieval. In *33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, pages 403–410. Association for Computing Machinery, 2010.

A. Ghose and P. Ipeirotis. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512, 2011.

P. Gil. Emoticons and Smileys 101, 2013. Available online, `http://netforbeginners.about.com/cs/netiquette101/a/bl_emoticons101.htm`.

A. Gliozzo and C. Strapparava. Cross Language Text Categorization by Acquiring Multilingual Domain Models from Comparable Corpora. In *ACL Workshop on Building and Using Parallel Texts (ParaText 2005)*, pages 9–16. Association for Computational Linguistics, 2005.

Google. Google Translate, 2007. Available online, `http://translate.google.com`.

I. Guyon. *Mining Massive Data Sets for Security: Advances in Data Mining, Search, Social Networks and Text Mining, and their Applications to Security*, chapter Practical Feature Selection: From Correlation to Causality, pages 27–43. IOS Press, 2008.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18, 2009.

V. Hatzivassiloglou and J. Wiebe. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In *18th International Conference on Computational Linguistics (COLING 2000)*, pages 299–305. Association for Computational Linguistics, 2000.

B. He, C. Macdonald, J. He, and I. Ounis. An Effective Statistical Approach to Blog Post Opinion Retrieval. In *17th ACM Conference on Information and Knowledge Management (CIKM 2008)*, pages 1063–1072. Association for Computing Machinery, 2008a.

B. He, C. Macdonald, and I. Ounis. Ranking Opinionated Blog Posts Using OpinionFinder. In *31st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 727–728. Association for Computing Machinery, 2008b.

B. Heerschop, F. Goossen, A. Hogenboom, F. Frasincar, U. Kaymak, and F. de Jong. Polarity Analysis of Texts using Discourse Structure. In *20th ACM Conference on Information and Knowledge Management (CIKM 2011)*, pages 1061–1070. Association for Computing Machinery, 2011a.

B. Heerschop, A. Hogenboom, and F. Frasincar. Sentiment Lexicon Creation from Lexical Resources. In *14th International Conference on Business Information Systems (BIS 2011)*, volume 87 of *Lecture Notes in Business Information Processing*, pages 185–196. Springer, 2011b.

B. Heerschop, P. van Iterson, A. Hogenboom, F. Frasincar, and U. Kaymak. Analyzing Sentiment in a Large Set of Web Data while Accounting for Negation. In *7th Atlantic Web Intelligence Conference (AWIC 2011)*, volume 86 of *Advances in Intelligent and Soft Computing*, pages 195–205. Springer, 2011c.

B. Heerschop, P. van Iterson, A. Hogenboom, F. Frasincar, and U. Kaymak. Accounting for Negation in Sentiment Analysis. In *11th Dutch-Belgian Information Retrieval Workshop (DIR 2011)*, pages 38–39, 2011d.

H. Hernault, H. Prendinger, D. duVerle, and M. Ishizuka. HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue and Discourse*, 1(3):1–33, 2010.

J. Herrera, A. Penas, and F. Verdejo. Techniques for Recognizing Textual Entailment and Semantic Equivalence. *Lecture Notes in Artificial Intelligence*, 4177:419–428, 2006.

K. Hofmann and V. Jijkoun. Generating a Non-English Subjectivity Lexicon: Relations that Matter. In *12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 398–405. Association for Computing Machinery, 2009.

A. Hogenboom, W. Ketter, J. van Dalen, U. Kaymak, J. Collins, and A. Gupta. Product Pricing using Adaptive Real-Time Probability of Acceptance Estimations based on

Economic Regimes. In *11th International Conference on Electronic Commerce (ICEC 2009)*, pages 176–185. Association for Computing Machinery, 2009a.

A. Hogenboom, V. Milea, F. Frasincar, and U. Kaymak. RCQ-GA: RDF Chain Query Optimization Using Genetic Algorithms. In *10th International Conference on Electronic Commerce and Web Technologies (EC-Web 2009)*, volume 5692 of *Lecture Notes in Computer Science*, pages 181–292. Springer, 2009b.

A. Hogenboom, V. Milea, F. Frasincar, and U. Kaymak. Genetic Algorithms for RDF Chain Query Optimization. In *21st Benelux Conference on Artificial Intelligence (BNAIC 2009)*, pages 327–328, 2009c.

A. Hogenboom, F. Hogenboom, U. Kaymak, W. Ketter, J. van Dalen, and J. Collins. Towards a Dynamic Model of Supply Chain Regimes for Complex Multi-Agent Markets. In *2010 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2010)*, pages 3219–3225. IEEE, 2010a.

A. Hogenboom, F. Hogenboom, U. Kaymak, P. Wouters, and F. de Jong. Mining Economic Sentiment using Argumentation Structures. In *7th International Workshop on Web Information Systems Modeling (WISM 2010) at the 29th International Conference on Conceptual Modeling (ER 2010)*, volume 6413 of *Lecture Notes in Computer Science*, pages 200–209. Springer, 2010b.

A. Hogenboom, V. Milea, F. Frasincar, and U. Kaymak. Optimizing RDF Chain Queries using Genetic Algorithms. In *Dutch-Belgian Database Day 2010 (DBDBD 2010)*, 2010c. Available online, `http://www.uhasselt.be/Documents/UHasselt/initiatieven/DBDBD-2010/HogenboomA.pdf`.

A. Hogenboom, P. van Iterson, B. Heerschop, F. Frasincar, and U. Kaymak. Determining Negation Scope and Strength in Sentiment Analysis. In *2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2011)*, pages 2589–2594. IEEE, 2011a.

A. Hogenboom, P. van Iterson, B. Heerschop, F. Frasincar, and U. Kaymak. Analyzing Sentiment while Accounting for Negation Scope and Strength. In *23rd Benelux Conference on Artificial Intelligence (BNAIC 2011)*, pages 327–328. Nevelland, 2011b.

A. Hogenboom, M. Bal, F. Frasincar, and D. Bal. Towards Cross-Language Sentiment Analysis through Universal Star Ratings. In *7th International Conference on Knowledge*

Management in Organizations (KMO 2012), volume 172 of *Advances in Intelligent Systems and Computing*, pages 69–79. Springer, 2012a.

A. Hogenboom, F. Boon, and F. Frasincar. A Statistical Approach to Star Rating Classification of Sentiment. In *1st International Symposium on Management Intelligent Systems (IS-MiS 2012)*, volume 171 of *Advances in Intelligent Systems and Computing*, pages 251–260. Springer, 2012b.

A. Hogenboom, M. Jongmans, and F. Frasincar. Structuring Political Documents for Importance Ranking. In *17th International Conference on Applications of Natural Language to Information Systems (NLDB 2012)*, volume 7337 of *Lecture Notes in Computer Science*, pages 345–350. Springer, 2012c.

A. Hogenboom, E. Niewenhuijse, F. Hogenboom, and F. Frasincar. RCQ-ACS: RDF Chain Query Optimization Using an Ant Colony System. In *2012 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2012)*, pages 74–81. IEEE, 2012d.

A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak. Exploiting Emoticons in Sentiment Analysis. In *28th Symposium on Applied Computing (SAC 2013)*, pages 703–710. Association for Computing Machinery, 2013a.

A. Hogenboom, F. Frasincar, and U. Kaymak. Ant Colony Optimization for RDF Chain Queries for Decision Support. *Expert Systems with Applications*, 40(5):1555–1563, 2013b.

A. Hogenboom, F. Hogenboom, F. Frasincar, K. Schouten, and O. van der Meer. Semantics-Based Information Extraction for Detecting Economic Events. *Multimedia Tools and Applications*, 64(1):27–52, 2013c.

A. Hogenboom, M. Bal, F. Frasincar, D. Bal, U. Kaymak, and F. de Jong. Lexicon-Based Sentiment Analysis by Mapping Conveyed Sentiment to Intended Sentiment. *International Journal of Web Engineering and Technology*, 9(2):125–147, 2014a.

A. Hogenboom, B. Heerschop, F. Frasincar, U. Kaymak, and F. de Jong. Multi-Lingual Support for Lexicon-Based Sentiment Analysis Guided by Semantics. *Decision Support Systems*, 66(1):43–53, 2014b.

A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak. Exploiting Emoticons in Polarity Classification of Text. *Journal of Web Engineering*, 14(1):22–40, 2015a.

A. Hogenboom, F. Frasincar, F. de Jong, and U. Kaymak. Using Rhetorical Structure in Sentiment Analysis. *Communications of the ACM*, 58(7):69–77, 2015b.

A. Hogenboom, F. Frasincar, F. de Jong, and U. Kaymak. Polarity Classification Using Structure-Based Vector Representations of Text. *Decision Support Systems*, 74(1):46–56, 2015c.

A. Hogenboom, W. Ketter, J. van Dalen, U. Kaymak, J. Collins, and A. Gupta. Adaptive Tactical Pricing in Multi-Agent Supply Chain Markets using Economic Regimes. *Decision Sciences Journal*, 46(4):791–818, 2015d.

A. Hogenboom, E. Niewenhuijse, M. Jansen, F. Frasincar, and D. Vandic. RDF Chain Query Optimization in a Distributed Environment. In *30th Symposium on Applied Computing (SAC 2015)*, pages 353–359. Association for Computing Machinery, 2015e.

F. Hogenboom, M. de Winter, F. Frasincar, and A. Hogenboom. A News-Based Approach for Computing Historical Value-at-Risk. In *1st International Symposium on Management Intelligent Systems (IS-MiS 2012)*, volume 171 of *Advances in Intelligent Systems and Computing*, pages 283–292. Springer, 2012e.

F. Hogenboom, M. de Winter, M. Jansen, A. Hogenboom, F. Frasincar, and U. Kaymak. Event-Based Historical Value-at-Risk. In *IEEE Computational Intelligence for Financial Engineering and Economics 2012 (CIFEr 2012)*, pages 164–170. IEEE, 2012f.

C. Holton. Identifying Disgruntled Employee Systems Fraud Risk Through Text Mining: A Simple Solution for a Multi-Billion Dollar Problem. *Decision Support Systems*, 46 (4):853–846, 2009.

E. Howrey. The Predictive Power of the Index of Consumer Sentiment. *Brookings Papers on Economic Activity*, 32(1):176–216, 2001.

M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pages 168–177. Association for Computing Machinery, 2004.

A. Ibrahim and T. Elghazaly. Rhetorical Representation and Vector Representation in Summarizing Arabic Text. In *18th International Conference on Applications of Natural Language to Information Systems (NLDB 2013)*, volume 7934 of *Lecture Notes in Computer Science*, pages 65–76. Springer, 2013.

Internet Live Stats. Google Search Statistics, 2014. Available online, `http://www.internetlivestats.com/google-search-statistics/`.

B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter Power: Tweets as Electronic Word of Mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188, 2009.

L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-Dependent Twitter Sentiment Classification. In *49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 151–160. Association for Computational Linguistics, 2011.

V. Jijkoun and K. Hofmann. Generating a Non-English Subjectivity Lexicon: Relations that Matter. In *12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 398–405. Association for Computational Linguistics, 2009.

D. Jurafsky and J. Martin. *Speech and Language Processing*. Prentice Hall, 2000.

J. Kamps, M. Marx, R. Mokken, and M. de Rijke. Using WordNet to Measure Semantic Orientations of Adjectives. In *3rd Conference on Language Resources and Evaluation (LREC 2004)*, pages 1115–1118. European Language Resources Association, 2004.

A. Kendon. On Gesture: Its Complementary Relationship with Speech. In *Nonverbal Communication*. Lawrence Erlbaum Associates, 1987.

A. Kennedy and D. Inkpen. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22(2):110–125, 2006.

S. Kim and E. Hovy. Determining the Sentiment of Opinions. In *20th International Conference on Computational Linguistics (COLING 2004)*, pages 1367–1373. Association for Computational Linguistics, 2004.

S. Kim and E. Hovy. Automatic Identification of Pro and Con Reasons in Online Reviews. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, pages 483–490. Association for Computational Linguistics, 2006.

Korte Reviews. Korte Reviews, 2011. Available online, `http://kortereviews.tumblr.com/`.

L. Kristoufek. BitCoin Meets Google Trends and Wikipedia: Quantifying the Relationship Between Phenomena of the Internet Era. *Scientific Reports*, 3(3415), 2013.

R. Krovetz. Viewing Morphology as an Inference Process. In *16th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993)*, pages 191–202. Association for Computing Machinery, 1993.

Y. Lee, S. Na, J. Kim, S. Nam, H. Jung, and J. Lee. KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval. In *17th Text Retrieval Conference (TREC 2008)*. National Institute of Standards and Technology, 2008.

J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6(2):167–195, 2015.

B. Lemaire. Lemaire Film Reviews, 2011. Available online, `http://www.lemairefilm.com/`.

K. Lerman, S. Blair-Goldensohn, and R. McDonald. Sentiment Summarization: Evaluating and Learning User Preferences. In *12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 514–522. Association for Computational Linguistics, 2009.

M. Lesk. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How To Tell a Pine Cone from an Ice Cream Cone. In *5th Annual International Conference on Systems Documentation (SIGDOC 1986)*, pages 24–26. Association for Computing Machinery, 1986.

E. Liddy. *Encyclopedia of Library and Information Science*, chapter Natural Language Processing, pages 2126–2136. Marcel Decker, 2nd edition, 2003.

Z. Lin, S. Tan, and X. Cheng. A Fast and Accurate Method for Bilingual Opinion Lexicon Extraction. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI 2012)*, pages 50–57. IEEE, 2012.

C. Lioma, B. Larsen, and W. Lu. Rhetorical Relations for Information Retrieval. In *35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012)*, pages 931–940. Association for Computing Machinery, 2012.

B. Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

K. Liu, W. Li, and M. Guo. Emoticon Smoothed Language Models for Twitter Sentiment Analysis. In *26th AAAI Conference on Artificial Intelligence (AAAI 2012)*, pages 1678–1684. Association for the Advancement of Artificial Intelligence, 2012.

S. Ludvigson. Consumer Confidence and Consumer Spending. *The Journal of Economic Perspectives*, 18(2):29–50, 2004.

A. Maas, R. Daly, P. Pham, D. Huang, A. Ng, and C. Potts. Learning Word Vectors for Sentiment Analysis. In *49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 142–150. Association for Computational Linguistics, 2011.

C. Macdonald and I. Ounis. The TREC Blogs 2006 Collection: Creating and Analysing a Blog Test Collection. Technical Report TR-2006-224, University of Glasgow, 2006. Available online, `http://terrierteam.dcs.gla.ac.uk/publications/macdonald06creating.pdf`.

S. Madden. From Databases to Big Data. *IEEE Internet Computing*, 16(3):4–6, 2012.

H. Mangassarian and H. Artail. A General Framework for Subjective Information Extraction from Unstructured English Text. *Data and Knowledge Engineering*, 62(2):352–367, 2007.

G. Mangnoesing, A. van Bunningen, A. Hogenboom, F. Hogenboom, and F. Frasincar. An Empirical Study for Determining Relevant Features for Sentiment Summarization of Online Conversational Documents. In *13th International Conference on Web Information Systems Engineering (WISE 2012)*, volume 7651 of *Lecture Notes in Computer Science*, pages 567–579. Springer, 2012.

W. Mann and S. Thompson. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281, 1988.

C. Manning, T. Grow, T. Grenager, J. Finkel, and J. Bauer. Stanford Tokenizer, 2010. Available online, `http://nlp.stanford.edu/software/tokenizer.shtml`.

C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 55–60. Association for Computational Linguistics, 2014.

Y. Mao and G. Lebanon. Sequential Models for Sentiment Prediction. In *ICML Workshop on Learning in Structured Output Spaces*, 2006.

D. Marcu. The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach. *Computational Linguistics*, 26(3):395–448, 2000.

T. Marks. Recommended Emoticons for Email Communication, 2004. Available online, `http://www.windweaver.com/emoticon.htm`.

J. Marshall. The Canonical Smiley (and 1-Line Symbol) List, 2003. Available online, `http://www.astro.umd.edu/~marshall/smileys.html`.

L. Marvin. Spoof, Spam, Lurk, and Lag: The Aesthetics of Text-Based Virtual Realities. *Journal of Computer-Mediated Communication*, 1(2), 1995.

Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. In *16th International Conference on World Wide Web (WWW 2007)*, pages 171–180. Association for Computing Machinery, 2007.

P. Melville, V. Sindhwani, and R. Lawrence. Social Media Analytics: Channeling the Power of the Blogosphere for Marketing Insight. In *1st Workshop on Information in Networks (WIN 2009)*, 2009.

Metacritic. Metacritic Reviews, 2011. Available online, `http://www.metacritic.com/browse/movies/title/dvd/`.

E. Metais. Enhancing Information Systems Management with Natural Language Processing Techniques. *Data and Knowledge Engineering*, 41(2):247–272, 2002.

R. Mihalcea, C. Banea, and J. Wiebe. Learning Multilingual Subjective Language via Cross-Lingual Projections. In *45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 976–983. Association for Computational Linguistics, 2007.

A. Montoyo, P. Martinez-Barco, and A. Balahur. Subjectivity and Sentiment Analysis: An Overview of the Current State of the Area and Envisaged Developments. *Decision Support Systems*, 53(4):675–679, 2012.

Msgweb. List of Emoticons in MSN Messenger, 2006. Available online, `http://www.msgweb.nl/en/MSN_Images/Emoticon_list/`.

T. Mullen and N. Collier. Sentiment Analysis Using Support Vector Machines with Diverse Information Sources. In *2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 412–418. Association for Computational Linguistics, 2004.

T. Mullen and R. Malouf. A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse. In *2006 AAAI Symposium on Computational Approaches to Analyzing Weblogs (AAAI-CAAW 2006)*, pages 159–162. Association for the Advancement of Artificial Intelligence, 2006.

R. Navigli. Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2): 1–69, 2009.

R. Navigli and P. Velardi. Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):671–674, 2005.

T. Nguyen and K. Shirai. Text Classification of Technical Papers Based on Text Segmentation. In *18th International Conference on Applications of Natural Language to Information Systems (NLDB 2013)*, volume 7934 of *Lecture Notes in Computer Science*, pages 278–284. Springer, 2013.

W. Nuij, V. Milea, F. Hogenboom, F. Frasincar, and U. Kaymak. An Automated Framework for Incorporating News into Stock Trading Strategies. *IEEE Transactions on Knowledge and Data Engineering*, 26(4):823–835, 2014.

N. O'Hare, M. Davy, A. Bermingham, P. Ferguson, P. Sheridan, C. Gurrin, and A. Smeaton. Topic-Dependent Sentiment Analysis of Financial Blogs. In *1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion (TSA 2009)*, pages 9–16. Association for Computing Machinery, 2009.

B. Ojokoh and O. Kayode. A Feature-Opinion Extraction Approach to Opinion Mining. *Journal of Web Engineering*, 11(1):51–63, 2012.

I. Ounis, C. Macdonald, and I. Soboroff. Overview of the TREC 2008 Blog Track. In *17th Text Retrieval Conference (TREC 2008)*. National Institute of Standards and Technology, 2008.

A. Pak and P. Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *7th Conference on International Language Resources and Evaluation (LREC 2010)*, pages 1320–1326. European Language Resources Association, 2010.

G. Paltoglou and M. Thelwall. A Study of Information Retrieval Weighting Schemes for Sentiment Analysis. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1386–1395. Association for Computational Linguistics, 2010.

B. Pang and L. Lee. A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts. In *42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 271–280. Association for Computational Linguistics, 2004.

B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1):1–135, 2008.

B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics, 2002.

J. Parapar, M. Vidal, and J. Santos. Finding the Best Parameter Setting: Particle Swarm Optimisation. In *2nd Spanish Conference on Information Retrieval (CERI 2012)*, pages 49–60. Springer, 2012.

Pingdom. Internet 2011 in Numbers, 2012. Available online, `http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers/`.

L. Polanyi and A. Zaenen. *Computing Attitude and Affect in Text: Theory and Applications*, chapter Contextual Valence Shifters, pages 1–10. Springer, 2006.

T. Preis, D. Reith, and H. Stanley. Complex Dynamics of our Economic Life on Different Scales: Insights from Search Engine Query Data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1933):5707–5719, 2010.

L. Qu, G. Ifrim, and G. Weikum. The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns. In *23rd International Conference on Computational Linguistics (COLING 2010)*, pages 913–921. Association for Computational Linguistics, 2010.

J. Read. Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In *Student Research Workshop at the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 43–48. Association for Computational Linguistics, 2005.

A. Reyes and P. Rosso. Making Objective Decisions from Subjective Data: Detecting Irony in Customer Reviews. *Decision Support Systems*, 53(4):754–760, 2012.

L. Rezabek and J. Cochenour. Visual Cues in Computer-Mediated Communication: Supplementing Text with Emoticons. *Journal of Visual Literacy*, 18(2):201–215, 1998.

S. Robertson. How Okapi came to TREC. In *TREC: Experiments and Evaluation in Information Retrieval*, pages 287–299. MIT Press, 2005.

S. Robertson and S. Walker. Okapi/Keenbow at TREC-8. In *8th Text Retrieval Conference (TREC-8)*, pages 151–162. National Institute of Standards and Technology, 2000.

H. Rui, Y. Liu, and A. Whinston. Whose and What Chatter Matters? The Effect of Tweets on Movie Sales. *Decision Support Systems*, 55(4):863–870, 2013.

R. Santos, B. He, C. Macdonald, and I. Ounis. Integrating Proximity to Subjective Sentences for Blog Opinion Retrieval. In *31st European Conference on Information Retrieval (ECIR 2009)*, pages 325–336. Springer, 2009.

R. Santos, C. Macdonald, R. McCreadie, I. Ounis, and I. Soboroff. Information Retrieval on the Blogosphere. *Foundations and Trends in Information Retrieval*, 6(1):1–125, 2012.

T. Scholz and S. Conrad. Extraction of Statements in News for a Media Response Analysis. In *18th International Conference on Applications of Natural Language to Information Systems (NLDB 2013)*, volume 7934 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2013a.

T. Scholz and S. Conrad. Linguistic Sentiment Features for Newspaper Opinion Mining. In *18th International Conference on Applications of Natural Language to Information Systems (NLDB 2013)*, volume 7934 of *Lecture Notes in Computer Science*, pages 272–277. Springer, 2013b.

K. Schouten and F. Frasincar. Finding Implicit Features in Consumer Reviews for Sentiment Analysis. In *14th International Conference on Web Engineering (ICWE 2014)*, volume 8541 of *Lecture Notes in Computer Science*, pages 130–144. Springer, 2014.

R. Schumaker, Y. Zhang, C. Huang, and H. Chen. Evaluating Sentiment in Financial News Articles. *Decision Support Systems*, 53(3):458–464, 2012.

Sharpened. Text-Based Emoticons, 2013. Available online, `http://www.sharpened.net/emoticons/`.

SharpNLP. SharpNLP, 2006. Available online, `http://sharpnlp.codeplex.com/`.

R. Shepard. Recognition Memory for Words, Sentences, and Pictures. *Journal of Verbal Learning and Verbal Behavior*, 6(1):156–163, 1967.

Short Reviews. Short Reviews, 2011. Available online, http://shortreviews.net/browse/.

T. Simpson and M. Crowe. WordNet.Net, 2005. Available online, http://opensource.ebswift.com/WordNet.Net.

R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1631–1642. Association for Computational Linguistics, 2013.

S. Somasundaran, G. Namata, L. Getoor, and J. Wiebe. Opinion Graphs for Polarity and Discourse Classification. In *2009 Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs 2009)*, pages 66–74. Association for Computational Linguistics, 2009a.

S. Somasundaran, G. Namata, J. Wiebe, and L. Getoor. Supervised and Unsupervised Methods in Employing Discourse Relations for Improving Opinion Polarity Classification. In *2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 170–179. Association for Computational Linguistics, 2009b.

R. Soricut and D. Marcu. Sentence Level Discourse Parsing using Syntactic and Lexical Information. In *Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL 2003)*, pages 149–156. Association for Computational Linguistics, 2003.

C. Strappavara and R. Mihalcea. SemEval-2007 Task 14: Affective Text. In *4th International Workshop on Semantic Evaluations (SemEval 2007)*, pages 70–74. Association for Computational Linguistics, 2007.

M. Taboada. Discourse Markers as Signals (or not) of Rhetorical Relations. *Journal of Pragmatics*, 38(4):567–592, 2006.

M. Taboada, K. Voll, and J. Brooke. Extracting Sentiment as a Function of Discourse Structure and Topicality. Technical Report 20, Simon Fraser University, 2008. Available online, http://www.cs.sfu.ca/research/publications/techreports/#2008.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307, 2011.

L. Tari, T. Phan, J. Hakenberg, Y. Chen, S. Tran, G. Gonzalez, and C. Baral. Incremental Information Extraction Using Relational Databases. *IEEE Transactions on Knowledge and Data Engineering*, 24(1):86–99, 2012.

M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment Strength Detection in Short Informal Text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

M. Thelwall, K. Buckley, and G. Paltoglou. SentiStrength, 2011. Available online, `http://sentistrength.wlv.ac.uk/`.

I. Titov and R. McDonald. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. pages 308–316, 2008.

P. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 417–424. Association for Computational Linguistics, 2002.

A. Tversky and D. Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, 1974.

A. van den Bosch, G. Busser, W. Daelemans, and S. Canisius. An Efficient Memory-Based Morphosyntactic Tagger and Parser for Dutch. In *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting (CLIN 2007)*, pages 93–114, 1997.

J. van der Meer, F. Boon, F. Hogenboom, F. Frasincar, and U. Kaymak. A Framework for Automatic Annotation of Web Pages Using the Google Rich Snippets Vocabulary. In *26th Symposium On Applied Computing (SAC 2011), Web Technologies Track*, pages 765–772. Association for Computing Machinery, 2011.

R. van Oest and P. Franses. Measuring Changes in Consumer Confidence. *Journal of Economic Psychology*, 29(3):255–275, 2008.

S. Vosen and T. Schmidt. Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends. *Journal of Forecasting*, 30(6):565–578, 2011.

P. Vossen. EuroWordNet: A Multilingual Database for Information Retrieval. In *3rd DELOS Workshop on Cross-Language Information Retrieval (DELOS 1997)*, pages 5–7. European Research Consortium for Informatics and Mathematics, 1997.

P. Vossen, K. Hofmann, M. de Rijke, E. Tjong, K. Sang, and K. Deschacht. The Cornetto Database: Architecture and User-Scenarios. In *7th Dutch-Belgian Information Retrieval Workshop (DIR 2007)*, pages 89–96. Acco, 2007.

J. Vuchelen. Consumer Sentiment and Macroeconomic Forecasts. *Journal of Economic Psychology*, 25(4):493–506, 2004.

B. Walenz and J. Didion. OpenNLP, 2008. Available online, `http://jwordnet.sourceforge.net/`.

X. Wan. Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In *2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 553–561. Association for Computational Linguistics, 2008.

X. Wan. Co-Training for Cross-Lingual Sentiment Classification. In *Joint Conference of the 47th Annual Meeting of ACL and the 4th International Join Conference on Natural Language Processing of the AFNLP (ACL 2009)*, pages 235–243. Association for Computational Linguistics, 2009.

X. Wang, X. Jin, M. Chen, K. Zhang, and D. Shen. Topic Mining over Asynchronous Text Sequences. *IEEE Transactions on Knowledge and Data Engineering*, 24(1):156–169, 2012.

C. Whitelaw, N. Garg, and S. Argamon. Using Appraisal Groups for Sentiment Analysis. In *14th ACM International Conference on Information and Knowledge Management (CIKM 2005)*, pages 625–631. Association for Computing Machinery, 2005.

J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning Subjective Language. *Computational Linguistics*, 30(3):277–308, 2004.

A. Wierzbicka. *Alternative Linguistics: Descriptive and Theoretical Modes*, chapter Dictionaries vs. Encyclopedias: How to Draw the Line, pages 289–316. John Benjamins Publishing Company, 1995a.

A. Wierzbicka. Emotion and Facial Expression: A Semantic Perspective. *Culture Psychology*, 1(2):227–258, 1995b.

Wikipedia. List of Emoticons, 2013. Available online, `http://en.wikipedia.org/wiki/List_of_emoticons/`.

Y. Wilks and M. Stevenson. The Grammar of Sense: Using Part-of-Speech Tags as a First Step in Semantic Disambiguation. *Journal of Natural Language Engineering*, 4 (2):135–143, 1998.

T. Wilson, J. Wiebe, and P. Hoffman. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 347–354. Association for Computational Linguistics, 2005.

D. Witmer and S. Katzman. On-Line Smiles: Does Gender Make a Difference in the Use of Graphic Accents? *Journal of Computer-Mediated Communication*, 2(4), 1997.

X. Yu, Y. Liu, X. Huang, and A. An. Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):720–734, 2012.

Y. Yu, W. Duan, and Q. Cao. The Impact of Social and Conventional Media on Firm Equity Value: A Sentiment Analysis Approach. *Decision Support Systems*, 55(4):919–926, 2013.

C. Zhai. Notes on the Lemur TFIDF Model. Technical report, Carnegie Mellon University, 2001. Available online, `http://www.lemurproject.org/lemur/tfidf.pdf`.

M. Zhang and X. Ye. A Generation Model to Unify Topic Relevance and Lexicon-Based Sentiment for Opinion Retrieval. In *31st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 411–418. Association for Computing Machinery, 2008.

B. Zhao, Z. Zhang, W. Qian, and A. Zhou. Identification of Collective Viewpoints on Microblogs. *Data and Knowledge Engineering*, 87(1):374–393, 2013.

J. Zhao, L. Dong, J. Wu, and K. Xu. MoodLens: An Emoticon-Based Sentiment Analysis System for Chinese Tweets. In *18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2012)*, pages 1528–1531. Association for Computing Machinery, 2012.

L. Zhou, B. Li, W. Gao, Z. Wei, and K. Wong. Unsupervised Discovery of Discourse Relations for Eliminating Intra-sentence Polarity Ambiguities. In *2011 Conference on*

Empirical Methods in Natural Language Processing (EMNLP 2011), pages 162–171. Association for Computational Linguistics, 2011.

C. Zirn, M. Niepert, H. Stuckenschmidt, and M. Strube. Fine-Grained Sentiment Analysis with Structural Features. In *5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 336–344. Asian Federation of Natural Language Processing, 2011.

Summary in English

Challenging economic conditions, speculative bubbles for cryptocurrencies like Bitcoin, and electronic word-of-mouth phenomena on social media like Twitter have recently demonstrated how today's markets are affected by people's sentiment. As moods and opinions play a pivotal role in various business and economic processes, keeping track of one's stakeholders' sentiment is crucial for today's decision makers.

Today's abundance and ubiquity of user-generated content allows for automated monitoring of the opinions of large quantities of (potential) stakeholders, such as consumers. Automated systems that perform such sentiment analysis tasks are mainly concerned with the extraction of subjective information from natural language text. One important challenge here lies in identifying whether pieces of text are positive or negative.

Typical methods of identifying this polarity of text involve low levels of linguistic analysis. Existing sentiment analysis systems predominantly use morphological, lexical, and syntactic cues for polarity, such as a text's words, the parts-of-speech of these words, and negation or amplification of the sentiment conveyed by some of the words (where applicable). However, the utilization of more, higher levels of linguistic analysis can improve a system's understanding of natural language. Therefore, the hypothesis underlying this dissertation is that the analysis of the polarity of text can be performed more accurately when additionally accounting for semantics and structure. The evaluation of this hypothesis has resulted in various findings.

First, the traditional morphological, lexical, and syntactic cues can help distinguish between polarity classes only to a limited extent. The way in which people express their sentiment may be context-dependent, thus rendering polarity classification solely based on low-level linguistic analysis particularly challenging.

Fortunately, some people use emoticons in natural language text in order to express, stress, or disambiguate their sentiment in particular text segments. Because of the interactions that emoticons have on a semantic level with other words in the text segments in which they occur, the sentiment associated with these emoticons tends to dominate the sentiment conveyed by the words, thus conveying valuable additional information.

Accounting for semantics turns out to be useful in a multi-lingual polarity classification setting as well. In this setting, machine-translating text to a reference language and subsequently analyzing the polarity of the translated text is an intuitive approach. Yet, this approach is a rather ineffective one, as machine translation may yield an inaccurate representation of the original content in the reference language, and semantics may be lost in translation. Conversely, a significantly better polarity classification performance can be achieved when exploiting semantic relations between and within languages in order to identify meaningful cues for sentiment in another language than the reference language.

An even better understanding of a text's conveyed sentiment can be obtained by performing linguistic analysis not only on a morphological, lexical, syntactic, and semantic level, but also on a discourse level. Automated sentiment analysis can be guided by a deep and fine-grained analysis of a text's rhetorical structure. A rule-based polarity classifier can thus exploit text segments' rhetorical roles in order to make a fine-grained distinction between important text segments and less important ones in terms of their contribution to conveying a text's overall sentiment, such that, e.g., conclusions can be treated differently from background information.

A drawback here is the computational complexity of the natural language processing techniques that automatically identify the rhetorical structure of a piece of text. However, an improved understanding of a carefully selected fraction of a text can already yield a significantly better understanding of the text as a whole, in terms of its conveyed sentiment. As such, the sentiment analysis process can be successfully guided by a shallow analysis of the rhetorical structure of only a small fraction of all text.

Even though rule-based sentiment analysis methods allow for intuitive ways of incorporating discourse analysis into the process, machine learning approaches can benefit from a more sophisticated linguistic analysis as well. Features representing the occurrence of specific (senses of) words are indispensable to good polarity classifiers. Nevertheless, these features can be successfully complemented by informative features that capture the sentiment conveyed by specific rhetorical elements that either constitute a text's core or provide crucial contextual information.

All in all, the findings of this dissertation suggest that the polarity of natural language text should not be determined solely based on *what* is said, but based on *how* the message of a text is conveyed as well. Promising applications of these findings lie in domains in which an accurate understanding of one's stakeholders' sentiment is warranted in order to enable effective support for decision making processes. Examples of such domains include marketing and reputation management.

Nederlandse Samenvatting (Summary in Dutch)

Stemmingen en meningen van mensen spelen vandaag de dag een belangrijke rol in uiteenlopende (bedrijfs)economische processen. Als ons de afgelopen tijd één ding duidelijk is gemaakt door de in zwaar weer verkerende economie, speculatieve zeepbellen rond cryptovaluta als Bitcoin, en mond-tot-mond fenomenen op sociale media als Twitter, dan is dat wel hoezeer de markten tegenwoordig beïnvloed kunnen worden door het sentiment van mensen. Het is dan ook van cruciaal belang om het sentiment van belanghebbenden mee te nemen in besluitvormingsprocessen.

De stortvloed aan informatie die dagelijks door talloze gebruikers on-line wordt gezet, maakt het mogelijk de meningen te monitoren van grote aantallen (potentiële) belanghebbenden, zoals consumenten. Geautomatiseerde systemen voor zulke sentimentanalyses zijn erop gericht subjectieve informatie uit tekst te halen. Een belangrijke uitdaging ligt hierbij in het bepalen of een tekst een positief of negatief sentiment uitdraagt.

Het bepalen van deze polariteit van tekst wordt doorgaans gedaan met behulp van taalkundige analyses op een laag niveau. Bestaande systemen voor geautomatiseerde sentimentanalyse gebruiken voornamelijk informatie op morfologisch, lexicaal, en syntactisch niveau om de polariteit van tekst te bepalen. Het gaat hierbij vaak om specifieke woorden, woordsoorten, en de eventuele negatie of amplificatie van het door bepaalde woorden uitgedragen sentiment. Tekst kan echter beter begrepen worden door meer, en vooral hogere niveaus van taalkundige analyse te benutten. De onderliggende hypothese van deze dissertatie is dan ook dat de polariteit van tekst nauwkeuriger bepaald kan worden door ook de semantiek en de structuur van tekst in de sentimentanalyse te betrekken. De evaluatie van deze hypothese heeft geresulteerd in diverse bevindingen.

Ten eerste helpt de traditionele morfologische, lexicale, en syntactische informatie slechts in beperkte mate polariteitsklasses te onderscheiden. De manier waarop sentiment tot uitdrukking komt, lijkt af te hangen van de context. Het lijkt dan een brug te ver om de polariteit van tekst enkel te classificeren via taalkundige analyses op een laag niveau.

Waardevolle extra informatie kan gehaald worden uit emoticons. Mensen gebruiken zulke emoticons om hun sentiment in specifieke tekstsegmenten uit te drukken, te benadrukken, of te verduidelijken. Emoticons hebben op een semantisch niveau interacties met andere woorden in de tekstsegmenten waarin ze voorkomen. Het sentiment geassocieerd met deze emoticons domineert daardoor het sentiment dat spreekt uit de woorden.

Rekening houden met semantiek blijkt ook nuttig te zijn bij het classificeren van de polariteit van teksten in verschillende talen. Een intuïtieve aanpak is om teksten automatisch naar één taal te vertalen, en vervolgens de polariteit van de vertalingen te bepalen. Dit is echter ineffectief, aangezien de originele tekst onnauwkeurig gerepresenteerd kan zijn in de vertaling, waar bovendien de semantiek deels verloren kan gaan. Het is daarentegen significant effectiever om semantische relaties tussen en binnen talen uit te buiten, om zo betekenisvolle aanwijzingen voor sentiment te identificeren in een andere taal.

Het sentiment dat een tekst uitdraagt kan nog beter worden begrepen door niet alleen taalkundige analyses uit te voeren op morfologisch, lexicaal, syntactisch, en semantisch niveau, maar ook op het niveau van discours. Zo kan sentimentanalyse geleid worden door een diepe, fijnmazige analyse van de retorische structuur van een tekst. Bij een op regels gebaseerde polariteitsclassificatie kan dan een subtiel onderscheid worden gemaakt tussen belangrijke en minder belangrijke tekstdelen, op basis van hun retorische rollen. Conclusies kunnen zo anders behandeld worden dan bijvoorbeeld achtergrondinformatie.

Een nadeel van deze aanpak is de computationele complexiteit van het automatisch herkennen van retorische structuren. Echter, qua uitgedragen sentiment kan een beter begrip van een zorgvuldig geselecteerd deel van een tekst al leiden tot een significant beter begrip van de volledige tekst. Sentimentanalyse kan dan ook geleid worden door een beperkte analyse van de retorische structuur van slechts een klein deel van alle tekst.

Discoursanalyse kan dus op een intuïtieve manier verwerkt worden in op regels gebaseerde sentimentanalyses. Automatisch lerende modellen voor polariteitsclassificatie kunnen echter ook baat hebben bij informatie uit dergelijke rijke taalkundige analyses. Voor zulke modellen is het onmisbaar om teksten met name te karakteriseren met behulp van specifieke (betekenissen van) woorden, maar het loont de moeite dit aan te vullen met een karakterisering van het sentiment dat wordt uitgedragen door specifieke retorische elementen, die de kern van een tekst vormen of cruciale contextuele informatie bevatten.

Al met al zou de polariteit van tekst niet enkel bepaald moeten worden op basis van *wat* men zegt, maar ook op basis van *hoe* een boodschap wordt overgebracht. Kansrijke toepassingen van deze bevindingen liggen in domeinen waar een nauwkeurig beeld van het sentiment van belanghebbenden cruciaal is voor een effectieve ondersteuning van besluitvormingsprocessen, bijvoorbeeld op het gebied van marketing of reputatiemanagement.

About the Author

Alexander Hogenboom (April 13, 1987) holds both a B.Sc. degree and a cum laude M.Sc. degree in Economics and Informatics, obtained at Erasmus University Rotterdam, The Netherlands, in 2007 and 2009, respectively. His research interests relate to the utilization of tools and techniques stemming from computer science in order to facilitate or support (business) economic processes. As such, Alexander's research covers semantic information systems, decision support systems, and intelligent systems for information extraction, with a specific focus on systems for automated sentiment analysis.

Since July 2009, Alexander has conducted his research in a Ph.D. candidacy under the auspices of the Erasmus Center for Business Intelligence (ECBI) at the Erasmus Research Institute of Management (ERIM), the Econometric Institute at the Erasmus School of Economics (ESE), Erasmus Studio, and the Dutch Research School for Information and Knowledge Systems (SIKS). Alexander's Ph.D. research is linked to the Argumentation Discovery in Economics Literature project of ERIM, the Semantic Scholarly Publishing project of Erasmus Studio, and the Infiniti project on Information Retrieval for Information Services (work package three) of the Dutch national program COMMIT.

Alexander has published 34 (peer-reviewed) papers in the proceedings of prestigious international conferences – e.g., BIS, CIKM, DEXA, ER, ICEC, NLDB, SAC, SMC, and WISE – and local venues like BNAIC, DBDBD, and DIR. His conference papers brought him an Honorable Mention Award (at ICEC 2009), as well as various travel grants. Moreover, Alexander has published 9 articles in renowned journals such as Communications of the ACM, Data and Knowledge Engineering, the Decision Sciences Journal, Decision Support Systems, Expert Systems with Applications, and the Journal of Web Engineering. Alexander is also a contributor to the EconomieOpinie and Backbone platforms.

In addition to his research activities, Alexander has acted as a reviewer for renowned journals like the Expert Systems with Applications journal and the Information Systems journal. Furthermore, he has been actively involved with international conferences, not only as participant, but also as a local organizer (e.g., for IFSA/EUSFLAT 2009 and DB-DBD 2013), session chair (e.g., for WISM 2010 and SMC 2011), and program committee member and reviewer (e.g., for SMC 2010 and SMC 2011).

Over the years, Alexander has played an active role at the Erasmus University Rotterdam as well. Between October 2009 and May 2013, he was a board member of the Erasmus Ph.D. Association Rotterdam (EPAR). In this position, he was responsible for internal and external communication. Furthermore, since July 2009, Alexander has been intensively involved with the supervision of 8 Bachelor's and 3 Master's theses. Additionally, he has been involved with coordinating and teaching many courses related to computer programming and information technology. Alexander's teaching activities have resulted in excellent student reviews, culminating in a nomination for a Professor of the Year Award for the first year of the International Business Administration (IBA) Bachelor's programme at the Rotterdam School of Management (RSM) in 2014.

ERIM Ph.D. Series Overview

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT (ERIM)

ERIM PH.D. SERIES RESEARCH IN MANAGEMENT

The ERIM PhD Series contains PhD dissertations in the field of Research in Management defended at Erasmus University Rotterdam and supervised by senior researchers affiliated to the `Erasmus Research Institute of Management (ERIM)`. All dissertations in the ERIM PhD Series are available in full text through the ERIM Electronic Series Portal: `http://repub.eur.nl/pub`. ERIM is the joint research institute of the Rotterdam School of Management (RSM) and the Erasmus School of Economics at the Erasmus University Rotterdam (EUR).

DISSERTATIONS LAST FIVE YEARS

Abbink, E.J., *Crew Management in Passenger Rail Transport*, Promotor(s): Prof.dr. L.G. Kroon & Prof.dr. A.P.M. Wagelmans, EPS-2014-325-LIS, `http://repub.eur.nl/pub/76927`

Acar, O.A., *Crowdsourcing for Innovation: Unpacking Motivational, Knowledge and Relational Mechanisms of Innovative Behavior in Crowdsourcing Platforms*, Promotor(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2014-321-LIS, `http://repub.eur.nl/pub/76076`

Acciaro, M., *Bundling Strategies in Global Supply Chains*, Promotor(s): Prof.dr. H.E. Haralambides, EPS-2010-197-LIS, `http://repub.eur.nl/pub/19742`

Akin Ates, M., *Purchasing and Supply Management at the Purchase Category Level: strategy, structure and performance*, Promotor(s): Prof.dr. J.Y.F. Wynstra & Dr. E.M. van Raaij, EPS-2014-300-LIS, `http://repub.eur.nl/pub/50283`

Akpinar, E., *Consumer Information Sharing*, Promotor(s): Prof.dr.ir. A. Smidts, EPS-2013-297-MKT, `http://repub.eur.nl/pub/50140`

Alexander, L., *People, Politics, and Innovation: A Process Perspective*, Promotor(s): Prof.dr. H.G. Barkema & Prof.dr. D.L. van Knippenberg, EPS-2014-331-S&E, http://repub.eur.nl/pub/77209

Alexiev, A.S., *Exploratory Innovation: The Role of Organizational and Top Management Team Social Capital*, Promotor(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-208-STR, http://repub.eur.nl/pub/20632

Almeida e Santos Nogueira, R.J. de, *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty*, Promotor(s): Prof.dr.ir. U. Kaymak & Prof.dr. J.M.C. Sousa, EPS-2014-310-LIS, http://repub.eur.nl/pub/51560

Bannouh, K., *Measuring and Forecasting Financial Market Volatility using High-Frequency Data*, Promotor(s): Prof.dr. D.J.C. van Dijk, EPS-2013-273-F&A, http://repub.eur.nl/pub/38240

Ben-Menahem, S.M., *Strategic Timing and Proactiveness of Organizations*, Promotor(s): Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2013-278-S&E, http://repub.eur.nl/pub/39128

Benning, T.M., *A Consumer Perspective on Flexibility in Health Care: Priority Access Pricing and Customized Care*, Promotor(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2011-241-MKT, http://repub.eur.nl/pub/23670

Berg, W.E. van den, *Understanding Salesforce Behavior using Genetic Association Studies*, Promotor(s): Prof.dr. W.J.M.I. Verbeke, EPS-2014-311-MKT, http://repub.eur.nl/pub/51440

Betancourt, N.E., *Typical Atypicality: Formal and Informal Institutional Conformity, Deviance, and Dynamics*, Promotor(s): Prof.dr. B. Krug, EPS-2012-262-ORG, http://repub.eur.nl/pub/32345

Bezemer, P.J., *Diffusion of Corporate Governance Beliefs: Board independence and the emergence of a shareholder value orientation in the Netherlands*, Promotor(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-192-STR, http://repub.eur.nl/pub/18458

Binken, J.L.G., *System markets: Indirect network effects in action, or inaction?*, Promotor(s): Prof.dr. S. Stremersch, EPS-2010-213-MKT, http://repub.eur.nl/pub/21186

Bliek, R. de, *Empirical Studies on the Economic Impact of Trust*, Promotor(s): Prof.dr. J. Veenman & Prof.dr. Ph.H.B.F. Franses, EPS-2015-324-ORG, http://repub.eur.nl/pub/78159

Blitz, D.C., *Benchmarking Benchmarks*, Promotor(s): Prof.dr. A.G.Z. Kemna & Prof.dr. W.F.C. Verschoor, EPS-2011-225-F&A, http://repub.eur.nl/pub/22624

Boons, M., *Working Together Alone in the Online Crowd: The Effects of Social Motivations and Individual Knowledge Backgrounds on the Participation and Performance of Members of Online Crowdsourcing Platforms*, Promotor(s): Prof.dr. H.G. Barkema & Dr. D.A. Stam, EPS-2014-306-S&E, http://repub.eur.nl/pub/50711

Borst, W.A.M., *Understanding Crowdsourcing: Effects of motivation and rewards on participation and performance in voluntary online activities*, Promotor(s): Prof.dr.ir. J.C.M. van den Ende & Prof.dr.ir. H.W.G.M. van Heck, EPS-2010-221-LIS, http://repub.eur.nl/pub/21914

Brazys, J., *Aggregated Marcoeconomic News and Price Discovery*, Promotor(s): Prof.dr. W.F.C. Verschoor, EPS-2015-351-F&A, http://repub.eur.nl/pub/78243

Budiono, D.P., *The Analysis of Mutual Fund Performance: Evidence from U.S. Equity Mutual Funds*, Promotor(s): Prof.dr. M.J.C.M. Verbeek & Dr.ir. M.P.E. Martens, EPS-2010-185-F&A, http://repub.eur.nl/pub/18126

Burger, M.J., *Structure and Cooptition in Urban Networks*, Promotor(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.R. Commandeur, EPS-2011-243-ORG, http://repub.eur.nl/pub/26178

Byington, E., *Exploring Coworker Relationships: Antecedents and Dimensions of Interpersonal Fit,Coworker Satisfaction, and Relational Models*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2013-292-ORG, http://repub.eur.nl/pub/41508

Camacho, N.M., *Health and Marketing: Essays on Physician and Patient Decision-Making*, Promotor(s): Prof.dr. S. Stremersch, EPS-2011-237-MKT, http://repub.eur.nl/pub/23604

Cancurtaran, P., *Essays on Accelerated Product Development*, Promotor(s): Prof.dr. F. Langerak & Prof.dr.ir. G.H. van Bruggen, EPS-2014-317-MKT, http://repub.eur.nl/pub/76074

Caron, E.A.M., *Explanation of Exceptional Values in Multi-dimensional Business Databases*, Promotor(s): Prof.dr.ir. H.A.M. Daniels & Prof.dr. G.W.J. Hendrikse, EPS-2013-296-LIS, `http://repub.eur.nl/pub/50005`

Carvalho, L. de, *Knowledge Locations in Cities: Emergence and Development Dynamics*, Promotor(s): Prof.dr. L. Berg, EPS-2013-274-S&E, `http://repub.eur.nl/pub/38449`

Carvalho de Mesquita Ferreira, L., *Attention Mosaics: Studies of Organizational Attention*, Promotor(s): Prof.dr. P.P.M.A.R. Heugens & Prof.dr. J. van Oosterhout, EPS-2010-205-ORG, `http://repub.eur.nl/pub/19882`

Cox, R.H.G.M., *To Own, To Finance, and To Insure - Residential Real Estate Revealed*, Promotor(s): Prof.dr. D. Brounen, EPS-2013-290-F&A, `http://repub.eur.nl/pub/40964`

Defilippi Angeldonis, E.F., *Access Regulation for Naturally Monopolistic Port Terminals: Lessons from Regulated Network Industries*, Promotor(s): Prof.dr. H.E. Haralambides, EPS-2010-204-LIS, `http://repub.eur.nl/pub/19881`

Deichmann, D., *Idea Management: Perspectives from Leadership, Learning, and Network Theory*, Promotor(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2012-255-ORG, `http://repub.eur.nl/pub/31174`

Deng, W., *Social Capital and Diversification of Cooperatives*, Promotor(s): Prof.dr. G.W.J. Hendrikse, EPS-2015-341-ORG, `http://repub.eur.nl/pub/77449`

Desmet, P.T.M., *In Money we Trust? Trust Repair and the Psychology of Financial Compensations*, Promotor(s): Prof.dr. D. de Cremer, EPS-2011-232-ORG, `http://repub.eur.nl/pub/23268`

Dietvorst, R.C., *Neural Mechanisms Underlying Social Intelligence and Their Relationship with the Performance of Sales Managers*, Promotor(s): Prof.dr. W.J.M.I. Verbeke, EPS-2010-215-MKT, `http://repub.eur.nl/pub/21188`

Dollevoet, T.A.B., *Delay Management and Dispatching in Railways*, Promotor(s): Prof.dr. A.P.M. Wagelmans, EPS-2013-272-LIS, `http://repub.eur.nl/pub/38241`

Doorn, S. van, *Managing Entrepreneurial Orientation*, Promotor(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch, & Prof.dr. H.W. Volberda, EPS-2012-258-STR, `http://repub.eur.nl/pub/32166`

Douwens-Zonneveld, M.G., *Animal Spirits and Extreme Confidence: No Guts, No Glory?*, Promotor(s): Prof.dr. W.F.C. Verschoor, EPS-2012-257-F&A, http://repub.eur.nl/pub/31914

Duca, E., *The Impact of Investor Demand on Security Offerings*, Promotor(s): Prof.dr. A. de Jong, EPS-2011-240-F&A, http://repub.eur.nl/pub/26041

Duursema, H., *Strategic Leadership: Moving Beyond the Leader-Follower Dyad*, Promotor(s): Prof.dr. R.J.M. van Tulder, EPS-2013-279-ORG, http://repub.eur.nl/pub/39129

Eck, N.J. van, *Methodological Advances in Bibliometric Mapping of Science*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2011-247-LIS, http://repub.eur.nl/pub/26509

Ellen, S. ter, *Measurement, Dynamics, and Implications of Heterogeneous Beliefs in Financial Markets*, Promotor(s): Prof.dr. W.F.C. Verschoor, EPS-2015-343-F&A, http://repub.eur.nl/pub/78191

Eskenazi, P.I., *The Accountable Animal*, Promotor(s): Prof.dr. F.G.H. Hartmann, EPS-2015-355-F&A, http://repub.eur.nl/pub/78300

Essen, M. van, *An Institution-Based View of Ownership*, Promotor(s): Prof.dr. J. van Oosterhout & Prof.dr. G.M.H. Mertens, EPS-2011-226-ORG, http://repub.eur.nl/pub/22643

Evangelidis, I., *Preference Construction under Prominence*, Promotor(s): Prof.dr. S.M.J. van Osselaer, EPS-2015-340-MKT, http://repub.eur.nl/pub/78202

Feng, L., *Motivation, Coordination and Cognition in Cooperatives*, Promotor(s): Prof.dr. G.W.J. Hendrikse, EPS-2010-220-ORG, http://repub.eur.nl/pub/21680

Fourne, S.P., *Managing Organizational Tensions: A Multi-Level Perspective on Exploration, Exploitation and Ambidexterity*, Promotor(s): Prof.dr. J.J.P. Jansen & Prof.dr. S.J. Magala, EPS-2014-318-S&E, http://repub.eur.nl/pub/76075

Gharehgozli, A.H., *Developing New Methods for Efficient Container Stacking Operations*, Promotor(s): Prof.dr.ir. M.B.M. de Koster, EPS-2012-269-LIS, http://repub.eur.nl/pub/37779

Gils, S. van, *Morality in Interactions: On the Display of Moral Behavior by Leaders and Employees*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2012-270-ORG, http://repub.eur.nl/pub/38027

Ginkel-Bieshaar, M.N.G. van, *The Impact of Abstract versus Concrete Product Communications on Consumer Decision-making Processes*, Promotor(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2012-256-MKT, http://repub.eur.nl/pub/31913

Gkougkousi, X., *Empirical Studies in Financial Accounting*, Promotor(s): Prof.dr. G.M.H. Mertens & Prof.dr. E. Peek, EPS-2012-264-F&A, http://repub.eur.nl/pub/37170

Glorie, K.M., *Clearing Barter Exchange Markets: Kidney Exchange and Beyond*, Promotor(s): Prof.dr. A.P.M. Wagelmans & Prof.dr. J.J. van de Klundert, EPS-2014-329-LIS, http://repub.eur.nl/pub/77183

Hakimi, N.A., *Leader Empowering Behaviour: The Leader's Perspective*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2010-184-ORG, http://repub.eur.nl/pub/17701

Hensmans, M., *A Republican Settlement Theory of the Firm: Applied to Retail Banks in England and the Netherlands (1830-2007)*, Promotor(s): Prof.dr. A. Jolink & Prof.dr. S.J. Magala, EPS-2010-193-ORG, http://repub.eur.nl/pub/19494

Hernandez-Mireles, C., *Marketing Modeling for New Products*, Promotor(s): Prof.dr. Ph.H.B.F. Franses, EPS-2010-202-MKT, http://repub.eur.nl/pub/19878

Heyde Fernandes, D. von der, *The Functions and Dysfunctions of Reminders*, Promotor(s): Prof.dr. S.M.J. van Osselaer, EPS-2013-295-MKT, http://repub.eur.nl/pub/41514

Heyden, M.L.M., *Essays on Upper Echelons & Strategic Renewal: A Multilevel Contingency Approach*, Promotor(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-259-STR, http://repub.eur.nl/pub/32167

Hoever, I.J., *Diversity and Creativity*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2012-267-ORG, http://repub.eur.nl/pub/37392

Hogenboom, F.P., *Automated Detection of Financial Events in News Text*, Promotor(s): Prof.dr.ir. U. Kaymak & Prof.dr. F.M.G. de Jong, EPS-2014-326-LIS, http://repub.eur.nl/pub/77237

Hoogendoorn, B., *Social Entrepreneurship in the Modern Economy: Warm Glow, Cold Feet*, Promotor(s): Prof.dr. H.P.G. Pennings & Prof.dr. A.R. Thurik, EPS-2011-246-STR, http://repub.eur.nl/pub/26447

Hoogervorst, N., *On The Psychology of Displaying Ethical Leadership: A Behavioral Ethics Approach*, Promotor(s): Prof.dr. D. de Cremer & Dr. M. van Dijke, EPS-2011-244-ORG, http://repub.eur.nl/pub/26228

Hout, D.H. van, *Measuring Meaningful Differences: Sensory Testing Based Decision Making in an Industrial Context; Applications of Signal Detection Theory and Thurstonian Modelling*, Promotor(s): Prof.dr. P.J.F. Groenen & Prof.dr. G.B. Dijksterhuis, EPS-2014-304-MKT, http://repub.eur.nl/pub/50387

Houwelingen, G.G. van, *Something To Rely On*, Promotor(s): Prof.dr. D. de Cremer & Prof.dr. M.H. van Dijke, EPS-2014-335-ORG, http://repub.eur.nl/pub/77320

Huang, X., *An Analysis of Occupational Pension Provision: From Evaluation to Redesign*, Promotor(s): Prof.dr. M.J.C.M. Verbeek, EPS-2010-196-F&A, http://repub.eur.nl/pub/19674

Hurk, E. van der, *Passengers, Information, and Disruptions*, Promotor(s): Prof.dr. L.G. Kroon & Prof.mr.dr. P.H.M. Vervest, EPS-2015-345-LIS, http://repub.eur.nl/pub/78275

Hytonen, K.A., *Context Effects in Valuation, Judgment and Choice: A Neuroscientific Approach*, Promotor(s): Prof.dr.ir. A. Smidts, EPS-2011-252-MKT, http://repub.eur.nl/pub/30668

Iseger, P. den, *Fourier and Laplace Transform Inversion with Applications in Finance*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2014-322-LIS, http://repub.eur.nl/pub/76954

Jaarsveld, W.L. van, *Maintenance Centered Service Parts Inventory Control*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2013-288-LIS, http://repub.eur.nl/pub/39933

Jalil, M.N., *Customer Information Driven After Sales Service Management: Lessons from Spare Parts Logistics*, Promotor(s): Prof.dr. L.G. Kroon, EPS-2011-222-LIS, http://repub.eur.nl/pub/22156

Kagie, M., *Advances in Online Shopping Interfaces: Product Catalog Maps and Recommender Systems*, Promotor(s): Prof.dr. P.J.F. Groenen, EPS-2010-195-MKT, http://repub.eur.nl/pub/19532

Kappe, E.R., *The Effectiveness of Pharmaceutical Marketing*, Promotor(s): Prof.dr. S. Stremersch, EPS-2011-239-MKT, http://repub.eur.nl/pub/23610

Karreman, B., *Financial Services and Emerging Markets*, Promotor(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.P.G. Pennings, EPS-2011-223-ORG, http://repub.eur.nl/pub/22280

Khanagha, S., *Dynamic Capabilities for Managing Emerging Technologies*, Promotor(s): Prof.dr. H.W. Volberda, EPS-2014-339-S&E, http://repub.eur.nl/pub/77319

Kil, J., *Acquisitions Through a Behavioral and Real Options Lens*, Promotor(s): Prof.dr. H.T.J. Smit, EPS-2013-298-F&A, http://repub.eur.nl/pub/50142

Klooster, E. van 't, *Travel to Learn: the Influence of Cultural Distance on Competence Development in Educational Travel*, Promotor(s): Prof.dr. F.M. Go & Prof.dr. P.J. van Baalen, EPS-2014-312-MKT, http://repub.eur.nl/pub/51462

Koendjbiharie, S.R., *The Information-Based View on Business Network Performance: Revealing the Performance of Interorganizational Networks*, Promotor(s): Prof.dr.ir. H.W.G.M. van Heck & Prof.mr.dr. P.H.M. Vervest, EPS-2014-315-LIS, http://repub.eur.nl/pub/51751

Koning, M., *The Financial Reporting Environment: The Role of the Media, Regulators and Auditors*, Promotor(s): Prof.dr. G.M.H. Mertens & Prof.dr. P.G.J. Roosenboom, EPS-2014-330-F&A, http://repub.eur.nl/pub/77154

Konter, D.J., *Crossing Borders with HRM: An Inquiry of the Influence of Contextual Differences in the Adoption and Effectiveness of HRM*, Promotor(s): Prof.dr. J. Paauwe & Dr. L.H. Hoeksema, EPS-2014-305-ORG, http://repub.eur.nl/pub/50388

Korkmaz, E., *Bridging Models and Business: Understanding Heterogeneity in Hidden Drivers of Customer Purchase Behavior*, Promotor(s): Prof.dr. S.L. van de Velde & Prof.dr. D. Fok, EPS-2014-316-LIS, http://repub.eur.nl/pub/76008

Kroezen, J.J., *The Renewal of Mature Industries: An Examination of the Revival of the Dutch Beer Brewing Industry*, Promotor(s): Prof.dr. P.P.M.A.R. Heugens, EPS-2014-333-S&E, http://repub.eur.nl/pub/77042

Kysucky, V., *Access to Finance in a Cros-Country Context*, Promotor(s): Prof.dr. L. Norden, EPS-2015-350-F&A, http://repub.eur.nl/pub/78225

Lam, K.Y., *Reliability and Rankings*, Promotor(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-230-MKT, http://repub.eur.nl/pub/22977

Lander, M.W., *Profits or Professionalism? On Designing Professional Service Firms*, Promotor(s): Prof.dr. J. van Oosterhout & Prof.dr. P.P.M.A.R. Heugens, EPS-2012-253-ORG, http://repub.eur.nl/pub/30682

Langhe, B. de, *Contingencies: Learning Numerical and Emotional Associations in an Uncertain World*, Promotor(s): Prof.dr.ir. B. Wierenga & Prof.dr. S.M.J. van Osselaer, EPS-2011-236-MKT, http://repub.eur.nl/pub/23504

Larco Martinelli, J.A., *Incorporating Worker-Specific Factors in Operations Management Models*, Promotor(s): Prof.dr.ir. J. Dul & Prof.dr.ir. M.B.M. de Koster, EPS-2010-217-LIS, http://repub.eur.nl/pub/21527

Leunissen, J.M., *All Apologies: On the Willingness of Perpetrators to Apologize*, Promotor(s): Prof.dr. D. de Cremer & Dr. M. van Dijke, EPS-2014-301-ORG, http://repub.eur.nl/pub/50318

Li, D., *Supply Chain Contracting for After-sales Service and Product Support*, Promotor(s): Prof.dr.ir. M.B.M. de Koster, EPS-2015-347-LIS, http://repub.eur.nl/pub/78526

Li, Z., *Irrationality: What, Why and How*, Promotor(s): Prof.dr. H. Bleichrodt, Prof.dr. P.P. Wakker, & Prof.dr. K.I.M. Rohde, EPS-2014-338-MKT, http://repub.eur.nl/pub/77205

Liang, Q.X., *Governance, CEO Identity, and Quality Provision of Farmer Cooperatives*, Promotor(s): Prof.dr. G.W.J. Hendrikse, EPS-2013-281-ORG, http://repub.eur.nl/pub/39253

Liket, K., *Why 'Doing Good' is not Good Enough: Essays on Social Impact Measurement*, Promotor(s): Prof.dr. H.R. Commandeur & Dr. K.E.H. Maas, EPS-2014-307-STR, http://repub.eur.nl/pub/51130

Loos, M.J.H.M. van der, *Molecular Genetics and Hormones: New Frontiers in Entrepreneurship Research*, Promotor(s): Prof.dr. A.R. Thurik, Prof.dr. P.J.F. Groenen, & Prof.dr. A. Hofman, EPS-2013-287-S&E, http://repub.eur.nl/pub/40081

Lovric, M., *Behavioral Finance and Agent-Based Artificial Markets*, Promotor(s): Prof.dr. J. Spronk & Prof.dr.ir. U. Kaymak, EPS-2011-229-F&A, http://repub.eur.nl/pub/22814

Lu, Y., *Data-Driven Decision Making in Auction Markets*, Promotor(s): Prof.dr.ir. H.W.G.M. van Heck & Prof.dr. W. Ketter, EPS-2014-314-LIS, http://repub.eur.nl/pub/51543

Manders, B., *Implementation and Impact of ISO 9001*, Promotor(s): Prof.dr. K. Blind, EPS-2014-337-LIS, http://repub.eur.nl/pub/77412

Markwat, T.D., *Extreme Dependence in Asset Markets Around the Globe*, Promotor(s): Prof.dr. D.J.C. van Dijk, EPS-2011-227-F&A, http://repub.eur.nl/pub/22744

Mees, H., *Changing Fortunes: How China's Boom Caused the Financial Crisis*, Promotor(s): Prof.dr. Ph.H.B.F. Franses, EPS-2012-266-MKT, http://repub.eur.nl/pub/34930

Meuer, J., *Configurations of Inter-firm Relations in Management Innovation: A Study in China's Biopharmaceutical Industry*, Promotor(s): Prof.dr. B. Krug, EPS-2011-228-ORG, http://repub.eur.nl/pub/22745

Micheli, M.R., *Business Model Innovation: A Journey across Managers' Attention and Inter-Organizational Networks*, Promotor(s): Prof.dr. J.J.P. Jansen, EPS-2015-344-S&E, http://repub.eur.nl/pub/78241

Mihalache, O.R., *Stimulating Firm Innovativeness: Probing the Interrelations between Managerial and Organizational Determinants*, Promotor(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch, & Prof.dr. H.W. Volberda, EPS-2012-260-S&E, http://repub.eur.nl/pub/32343

Milea, V., *News Analytics for Financial Decision Support*, Promotor(s): Prof.dr.ir. U. Kaymak, EPS-2013-275-LIS, http://repub.eur.nl/pub/38673

Naumovska, I., *Socially Situated Financial Markets: A Neo-Behavioral Perspective on Firms, Investors and Practices*, Promotor(s): Prof.dr. P.P.M.A.R. Heugens & Prof.dr. A. de Jong, EPS-2014-319-S&E, http://repub.eur.nl/pub/76084

Nielsen, L.K., *Rolling Stock Rescheduling in Passenger Railways: Applications in short-term planning and in disruption management*, Promotor(s): Prof.dr. L.G. Kroon, EPS-2011-224-LIS, http://repub.eur.nl/pub/22444

Nijdam, M.H., *Leader Firms: The value of companies for the competitiveness of the Rotterdam seaport cluster*, Promotor(s): Prof.dr. R.J.M. van Tulder, EPS-2010-216-ORG, http://repub.eur.nl/pub/21405

Noordegraaf-Eelens, L.H.J., *Contested Communication; A Critical Analysis of Central Bank Speech,* Promotor(s): Prof.dr. Ph.H.B.F. Franses, Prof.dr. J. de Mul, & Prof.dr. D.J.C. van Dijk, EPS-2010-209-MKT, http://repub.eur.nl/pub/21061

Nuijten, A.L.P., *Deaf Effect for Risk Warnings: A Causal Examination applied to Information Systems Projects*, Promotor(s): Prof.dr. G.J. van der Pijl, Prof.dr. H.R. Commandeur, & Prof.dr. M. Keil, EPS-2012-263-S&E, http://repub.eur.nl/pub/34928

Oosterhout, M. van, *Business Agility and Information Technology in Service Organizations*, Promotor(s): Prof.dr.ir. H.W.G.M. van Heck, EPS-2010-198-LIS, http://repub.eur.nl/pub/19805

Osadchiy, S.E., *The Dynamics of Formal Organization: Essays on bureaucracy and formal rules*, Promotor(s): Prof.dr. P.P.M.A.R. Heugens, EPS-2011-231-ORG, http://repub.eur.nl/pub/23250

Otgaar, A.H.J., *Industrial Tourism: Where the Public Meets the Private*, Promotor(s): Prof.dr. L. Berg, EPS-2010-219-ORG, http://repub.eur.nl/pub/21585

Ozdemir, M.N., *Project-level Governance, Monetary Incentives, and Performance in Strategic R&D Alliances*, Promotor(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2011-235-LIS, http://repub.eur.nl/pub/23550

Peers, Y., *Econometric Advances in Diffusion Models*, Promotor(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-251-MKT, http://repub.eur.nl/pub/30586

Peters, M., *Machine Learning Algorithms for Smart Electricity Markets*, Promotor(s): Prof.dr. W. Ketter, EPS-2014-332-LIS, http://repub.eur.nl/pub/77413

Pince, C., *Advances in Inventory Management: Dynamic Models*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2010-199-LIS, http://repub.eur.nl/pub/19867

Porck, J., *No Team is an Island: An Integrative View of Strategic Consensus between Groups*, Promotor(s): Prof.dr. P.J.F. Groenen & Prof.dr. D.L. van Knippenberg, EPS-2013-299-ORG, http://repub.eur.nl/pub/50141

Porras Prado, M., *The Long and Short Side of Real Estate, Real Estate Stocks, and Equity*, Promotor(s): Prof.dr. M.J.C.M. Verbeek, EPS-2012-254-F&A, http://repub.eur.nl/pub/30848

Poruthiyil, P.V., *Steering Through: How organizations negotiate permanent uncertainty and unresolvable choices*, Promotor(s): Prof.dr. P.P.M.A.R. Heugens & Prof.dr. S.J. Magala, EPS-2011-245-ORG, `http://repub.eur.nl/pub/26392`

Potthoff, D., *Railway Crew Rescheduling: Novel approaches and extensions*, Promotor(s): Prof.dr. A.P.M. Wagelmans & Prof.dr. L.G. Kroon, EPS-2010-210-LIS, `http://repub.eur.nl/pub/21084`

Pourakbar, M., *End-of-Life Inventory Decisions of Service Parts*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2011-249-LIS, `http://repub.eur.nl/pub/30584`

Pronker, E.S., *Innovation Paradox in Vaccine Target Selection*, Promotor(s): Prof.dr. H.J.H.M. Claassen & Prof.dr. H.R. Commandeur, EPS-2013-282-S&E, `http://repub.eur.nl/pub/39654`

Pruijssers, J.K., *An Organizational Perspective on Auditor Conduct*, Promotor(s): Prof.dr. J. van Oosterhout & Prof.dr. P.P.M.A.R. Heugens, EPS-2015-342-S&E, `http://repub.eur.nl/pub/78192`

Retel Helmrich, M.J., *Green Lot-Sizing*, Promotor(s): Prof.dr. A.P.M. Wagelmans, EPS-2013-291-LIS, `http://repub.eur.nl/pub/41330`

Rietveld, N., *Essays on the Intersection of Economics and Biology*, Promotor(s): Prof.dr. A.R. Thurik, Prof.dr. Ph.D. Koellinger, Prof.dr. P.J.F. Groenen, & Prof.dr. A. Hofman, EPS-2014-320-S&E, `http://repub.eur.nl/pub/76907`

Rijsenbilt, J.A., *CEO Narcissism: Measurement and Impact*, Promotor(s): Prof.dr. A.G.Z. Kemna & Prof.dr. H.R. Commandeur, EPS-2011-238-STR, `http://repub.eur.nl/pub/23554`

Roelofsen, E.M., *The Role of Analyst Conference Calls in Capital Markets*, Promotor(s): Prof.dr. G.M.H. Mertens & Prof.dr. L.G. van der Tas, EPS-2010-190-F&A, `http://repub.eur.nl/pub/18013`

Roza-van Vuren, M.W., *The Relationship between Offshoring Strategies and Firm Performance: Impact of innovation, absorptive capacity and firm size*, Promotor(s): Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2011-214-STR, `http://repub.eur.nl/pub/22155`

Rubbaniy, G., *Investment Behaviour of Institutional Investors*, Promotor(s): Prof.dr. W.F.C. Verschoor, EPS-2013-284-F&A, `http://repub.eur.nl/pub/40068`

Schellekens, G.A.C., *Language Abstraction in Word of Mouth*, Promotor(s): Prof.dr.ir. A. Smidts, EPS-2010-218-MKT, `http://repub.eur.nl/pub/21580`

Shahzad, K., *Credit Rating Agencies, Financial Regulations and the Capital Markets*, Promotor(s): Prof.dr. G.M.H. Mertens, EPS-2013-283-F&A, `http://repub.eur.nl/pub/39655`

Sotgiu, F., *Not All Promotions are Made Equal: From the Effects of a Price War to Cross-chain Cannibalization*, Promotor(s): Prof.dr. M.G. Dekimpe & Prof.dr.ir. B. Wierenga, EPS-2010-203-MKT, `http://repub.eur.nl/pub/19714`

Sousa, M.J.C. de, *Servant Leadership to the Test: New Perspectives and Insights*, Promotor(s): Prof.dr. D.L. van Knippenberg & Dr. D. van Dierendonck, EPS-2014-313-ORG, `http://repub.eur.nl/pub/51537`

Spliet, R., *Vehicle Routing with Uncertain Demand*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2013-293-LIS, `http://repub.eur.nl/pub/41513`

Srour, F.J., *Dissecting Drayage: An Examination of Structure, Information, and Control in Drayage Operations*, Promotor(s): Prof.dr. S.L. van de Velde, EPS-2010-186-LIS, `http://repub.eur.nl/pub/18231`

Staadt, J.L., *Leading Public Housing Organisation in a Problematic Situation: A Critical Soft Systems Methodology Approach*, Promotor(s): Prof.dr. S.J. Magala, EPS-2014-308-ORG, `http://repub.eur.nl/pub/50712`

Stallen, M., *Social Context Effects on Decision-Making: A Neurobiological Approach*, Promotor(s): Prof.dr.ir. A. Smidts, EPS-2013-285-MKT, `http://repub.eur.nl/pub/39931`

Tarakci, M., *Behavioral Strategy: Strategic Consensus, Power and Networks*, Promotor(s): Prof.dr. D.L. van Knippenberg & Prof.dr. P.J.F. Groenen, EPS-2013-280-ORG, `http://repub.eur.nl/pub/39130`

Teixeira de Vasconcelos, M., *Agency Costs, Firm Value, and Corporate Investment*, Promotor(s): Prof.dr. P.G.J. Roosenboom, EPS-2012-265-F&A, `http://repub.eur.nl/pub/37265`

Tempelaar, M.P., *Organizing for Ambidexterity: Studies on the pursuit of exploration and exploitation through differentiation, integration, contextual and individual attributes*, Promotor(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2010-191-STR, http://repub.eur.nl/pub/18457

Tiwari, V., *Transition Process and Performance in IT Outsourcing: Evidence from a Field Study and Laboratory Experiments*, Promotor(s): Prof.dr.ir. H.W.G.M. van Heck & Prof.mr.dr. P.H.M. Vervest, EPS-2010-201-LIS, http://repub.eur.nl/pub/19868

Troster, C., *Nationality Heterogeneity and Interpersonal Relationships at Work*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2011-233-ORG, http://repub.eur.nl/pub/23298

Tsekouras, D., *No Pain No Gain: The Beneficial Role of Consumer Effort in Decision-Making*, Promotor(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2012-268-MKT, http://repub.eur.nl/pub/37542

Tuijl, E. van, *Upgrading across Organisational and Geographical Configurations*, Promotor(s): Prof.dr. L. van den Berg, EPS-2015-349-S&E, http://repub.eur.nl/pub/78224

Tuncdogan, A., *Decision Making and Behavioral Strategy: The Role of Regulatory Focus in Corporate Innovation Processes*, Promotor(s): Prof.dr.ing. F.A.J. van den Bosch, Prof.dr. H.W. Volberda, & Prof.dr. T.J.M. Mom, EPS-2014-334-S&E, http://repub.eur.nl/pub/76978

Tzioti, S., *Let Me Give You a Piece of Advice: Empirical Papers about Advice Taking in Marketing*, Promotor(s): Prof.dr. S.M.J. van Osselaer & Prof.dr.ir. B. Wierenga, EPS-2010-211-MKT, http://repub.eur.nl/pub/21149

Uijl, S. den, *The Emergence of De-facto Standards*, Promotor(s): Prof.dr. K. Blind, EPS-2014-328-LIS, http://repub.eur.nl/pub/77382

Vaccaro, I.G., *Management Innovation: Studies on the Role of Internal Change Agents*, Promotor(s): Prof.dr.ing. F.A.J. van den Bosch, Prof.dr. H.W. Volberda, & Prof.dr. J.J.P. Jansen, EPS-2010-212-STR, http://repub.eur.nl/pub/21150

Vagias, D., *Liquidity, Investors and International Capital Markets*, Promotor(s): Prof.dr. M.A. van Dijk, EPS-2013-294-F&A, http://repub.eur.nl/pub/41511

Veelenturf, L.P., *Disruption Management in Passenger Railways: Models for Timetable, Rolling Stock and Crew Rescheduling*, Promotor(s): Prof.dr. L.G. Kroon, EPS-2014-327-LIS, http://repub.eur.nl/pub/77155

Venus, M., *Demystifying Visionary Leadership: In search of the essence of effective vision communication*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2013-289-ORG, http://repub.eur.nl/pub/40079

Verheijen, H.J.J., *Vendor-Buyer Coordination in Supply Chains*, Promotor(s): Prof.dr.ir. J.A.E.E. van Nunen, EPS-2010-194-LIS, http://repub.eur.nl/pub/19594

Visser, V.A., *Leader Affect and Leadership Effectiveness:How leader affective displays influence follower outcomes*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2013-286-ORG, http://repub.eur.nl/pub/40076

Vlam, A.J., *Customer First? The Relationship between Advisors and Consumers of Financial Products*, Promotor(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-250-MKT, http://repub.eur.nl/pub/30585

Waard, E.J. de, *Engaging Environmental Turbulence: Organizational Determinants for Repetitive Quick and Adequate Responses*, Promotor(s): Prof.dr. H.W. Volberda & Prof.dr. J. Soeters, EPS-2010-189-STR, http://repub.eur.nl/pub/18012

Waltman, L., *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*, Promotor(s): Prof.dr.ir. R. Dekker & Prof.dr.ir. U. Kaymak, EPS-2011-248-LIS, http://repub.eur.nl/pub/26564

Wang, T., *Essays in Banking and Corporate Finance*, Promotor(s): Prof.dr. L. Norden & Prof.dr. P.G.J. Roosenboom, EPS-2015-352-F&A, http://repub.eur.nl/pub/78301

Wang, Y., *Information Content of Mutual Fund Portfolio Disclosure*, Promotor(s): Prof.dr. M.J.C.M. Verbeek, EPS-2011-242-F&A, http://repub.eur.nl/pub/26066

Wang, Y., *Corporate Reputation Management: Reaching Out to Financial Stakeholders*, Promotor(s): Prof.dr. C.B.M. van Riel, EPS-2013-271-ORG, http://repub.eur.nl/pub/38675

Weenen, T.C., *On the Origin and Development of the Medical Nutrition Industry*, Promotor(s): Prof.dr. H.R. Commandeur & Prof.dr. H.J.H.M. Claassen, EPS-2014-309-S&E, http://repub.eur.nl/pub/51134

Wolfswinkel, M., *Corporate Governance, Firm Risk and Shareholder Value*, Promotor(s): Prof.dr. A. de Jong, EPS-2013-277-F&A, http://repub.eur.nl/pub/39127

Wubben, M.J.J., *Social Functions of Emotions in Social Dilemmas*, Promotor(s): Prof.dr. D. de Cremer & Prof.dr. E. van Dijk, EPS-2010-187-ORG, http://repub.eur.nl/pub/18228

Xu, Y., *Empirical Essays on the Stock Returns, Risk Management, and Liquidity Creation of Banks*, Promotor(s): Prof.dr. M.J.C.M. Verbeek, EPS-2010-188-F&A, http://repub.eur.nl/pub/18125

Yang, S., *Information Aggregation Efficiency of Prediction Markets*, Promotor(s): Prof.dr.ir. H.W.G.M. van Heck, EPS-2014-323-LIS, http://repub.eur.nl/pub/77184

Zaerpour, N., *Efficient Management of Compact Storage Systems*, Promotor(s): Prof.dr.ir. M.B.M. de Koster, EPS-2013-276-LIS, http://repub.eur.nl/pub/38766

Zhang, D., *Essays in Executive Compensation*, Promotor(s): Prof.dr. I. Dittmann, EPS-2012-261-F&A, http://repub.eur.nl/pub/32344

Zhang, X., *Scheduling with Time Lags*, Promotor(s): Prof.dr. S.L. van de Velde, EPS-2010-206-LIS, http://repub.eur.nl/pub/19928

Zhou, H., *Knowledge, Entrepreneurship and Performance: Evidence from country-level and firm-level studies*, Promotor(s): Prof.dr. A.R. Thurik & Prof.dr. L.M. Uhlaner, EPS-2010-207-ORG, http://repub.eur.nl/pub/20634

Zwan, P.W. van der, *The Entrepreneurial Process: An International Analysis of Entry and Exit*, Promotor(s): Prof.dr. A.R. Thurik & Prof.dr. P.J.F. Groenen, EPS-2011-234-ORG, http://repub.eur.nl/pub/23422