Lexicon-Based Sentiment Analysis by Mapping Conveyed Sentiment to Intended Sentiment

Alexander Hogenboom*
Malissa Bal
Flavius Frasincar
Daniella Bal

Econometric Institute
Erasmus University Rotterdam
P.O. Box 1738, NL-3000 DR Rotterdam, the Netherlands
E-mail: {hogenboom, frasincar}@ese.eur.nl,
{malissa.bal, daniella.bal}@xs4all.nl
Fax: +31 (0)10 408 9162
*Corresponding author

Uzay Kaymak

Industrial Engineering & Innovation Sciences
Eindhoven University of Technology
P.O. Box 513, NL-5600 MB Eindhoven, the Netherlands
E-mail: u.kaymak@ieee.org
Fax: +31 (0)40 243 2612

Franciska de Jong

Department of Computer Science
Universiteit Twente
P.O. Box 217, NL-7500 AE Enschede, the Netherlands
E-mail: f.m.g.dejong@utwente.nl
Fax: +31 (0)53 489 3503

Abstract: As consumers nowadays generate increasingly more content describing their experiences with, e.g., products and brands in various languages, information systems monitoring 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 sentiment conveyed by unstructured natural language text to universal star ratings, capturing intended sentiment. For these mappings, we consider a monotonically increasing step function, a naïve Bayes method, and a support vector machine. We demonstrate that the way in which natural language reveals intended sentiment differs across our data sets of Dutch and English texts. Additionally, the results of our experiments on modelling the relation between conveyed sentiment and intended sentiment suggest that language-specific sentiment scores can separate universal classes of intended sentiment from one another to a limited extent.

Keywords: Sentiment analysis; Star ratings; Sentiment mappings; Naïve Bayes; Support vector machine.

Reference to this paper should be made as follows: Hogenboom, A., Bal, M., Frasincar, F., Bal, D., Kaymak, U., and De Jong, F. (xxxx) 'Lexicon-Based Sentiment Analysis by Mapping Conveyed Sentiment to Intended Sentiment', *Int. J. Web Engineering and Technology*, Vol. x, No. x, pp. xxx–xxx.

Biographical notes: Alexander Hogenboom holds 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. Alexander is currently a Ph.D. candidate at Erasmus University Rotterdam. He is affiliated with the Econometric Institute, the research centre for Business Intelligence at the Erasmus Research Institute of Management, and Erasmus Studio. In his research, Alexander explores the utilisation of methods and techniques from informatics for facilitating and supporting decision making processes. His research interests relate to intelligent systems for information extraction, focused on tracking and monitoring of economic sentiment.

Malissa Bal obtained a B.Sc. degree in economics and informatics at Erasmus University Rotterdam, the Netherlands, in 2011 and is currently pursuing an M.Sc. degree in business information management at the same university. Her research interests range from multi-lingual sentiment analysis to the use of social media in business-to-consumer and business-to-business environments in general.

Flavius Frasincar obtained the M.Sc. degree in computer science from Politehnica University Bucharest, Romania, in 1998. In 2000, he received the professional doctorate degree in software engineering from Eindhoven University of Technology, the Netherlands. He got the Ph.D. degree in computer science from Eindhoven University of Technology, the Netherlands, in 2005. Since 2005, he is assistant professor in information systems at the Econometric Institute of Erasmus University Rotterdam, the Netherlands. He has published in numerous conferences and journals in the areas of databases, Web information systems, personalisation, and the Semantic Web. He is a member of the editorial board of the International Journal of Web Engineering and Technology.

Daniella Bal holds a B.Sc. degree in economics informatics, which she obtained at Erasmus University Rotterdam, the Netherlands, in 2011. Daniella is currently working towards her M.Sc. degree in business information management at Erasmus University Rotterdam. Her research interests cover various aspects of sentiment analysis, ranging from the exploitation of emoticons in the sentiment analysis process to sentiment analysis in a multi-lingual setting.

Uzay Kaymak received the M.Sc. degree in electrical engineering, the degree of chartered designer in information technology, and the Ph.D. degree in control engineering from the Delft University of Technology, Delft, the Netherlands, in 1992, 1995, and 1998, respectively. From 1997 to 2000, he was a reservoir engineer with Shell International Exploration and Production. Currently, he is full professor of information systems in health care at the department of Industrial Engineering & Innovation Sciences of the Eindhoven University of Technology, the Netherlands. Prof. Kaymak is an associate editor of IEEE Transactions on Fuzzy Systems and is a member of the editorial board of several journals.

Franciska de Jong is full professor of language technology at the University of Twente, the Netherlands, since 1992. She is also affiliated with the Erasmus University Rotterdam, the Netherlands, where she is director of the Erasmus Studio. She studied Dutch language and literature at the university of Utrecht, the Netherlands, did a Ph.D. track in theoretical linguistics and started to work on language technology in 1985 at Philips Research, where she focused on machine translation. Currently, her main research interest is in the field of multimedia indexing, text mining, semantic access, cross-language retrieval, and the disclosure of cultural heritage collections (in particular spoken audio archives). She is frequently involved in international programme committees, expert groups, and review panels, and has initiated a number of European Union projects. Since 2008, she is a member of the Governing Board of the Netherlands Organisation for Scientific Research (NWO).

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. Recent estimates indicate that 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 & 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 & Ahmad, 2007; Ding et al., 2008; Heerschop et al., 2011*a*,*b*; Taboada et al., 2011; Hogenboom et al., 2012), not in the least because 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., 2013) – 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., 2011*a*).

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. These scores are typically defined on a continuous scale between, e.g., −1 (negative) and 1 (positive). Intuitively, such scores should be meaningful and comparable across languages, irrespective of the methods used to obtain these scores. Therefore, many existing methods of analyzing sentiment in a multi-lingual setting focus on language-specific sentiment analysis, and subsequently treat all language-specific sentiment scores equally in a cross-lingual analysis (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, 1995*a*,*b*).

The language-specific sentiment scores produced by existing sentiment analysis methods typically reflect the sentiment conveyed by 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 analysing 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 our current endeavours, we aim to gain insight in the relation between language-specific scores of conveyed sentiment 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 paper is structured as follows. First, we discuss related work on sentiment analysis in Section 2. Then, in Section 3, we propose several methods for mapping language-specific sentiment scores to universal classifications of intended sentiment in order to facilitate more meaningful analyses of people's true sentiment. A discussion of insights following from the evaluation of our methods on Dutch and English documents is presented in Section 4. Last, we conclude and propose directions for future work in Section 5.

2 Related Work

In an extensive literature survey on sentiment analysis (Pang & Lee, 2008), the current surge of research interest in systems that deal with opinions and sentiment is attributed to the fact that, despite today's users' hunger for and reliance upon on-line advice and recommendations, explicit information on user opinions is often hard to find, confusing, or overwhelming. Many sentiment analysis approaches exist (Cambria et al., 2013; Feldman, 2013), yet the relation between language-specific sentiment scores stemming from such approaches and the actually intended sentiment has been relatively unexplored.

2.1 Sentiment Analysis

The main objective of most sentiment analysis methods is to extract subjective information from natural language text. Existing work focuses on several specific tasks. Some work aims to distinguish subjective text segments from objective text segments or to identify the degree of subjectivity of text (Wiebe et al., 2004). Other work focuses on determining the overall polarity of words, sentences, text segments, or documents (Pang & Lee, 2008). Even though this task is commonly approached as a binary classification problem, in which a text is to be classified as either positive or negative, some research focuses on ternary classification by introducing a third class of neutral documents. Other work even focuses on determining the degree of positivity or negativity of natural language text in order to produce, e.g., rankings of positive and negative documents (Chenlo & Losada, 2011; Chenlo et al., 2013).

There are two main types of approaches to sentiment classification tasks. On the one hand, some approaches exploit (generic) sentiment lexicons when determining the subjectivity or polarity of natural language text. On the other hand, many state-of-the-art approaches rely on machine learning techniques for sentiment analysis. Recently, hybrid approaches that combine machine learning techniques with sentiment lexicons have emerged as well.

Lexicon-based sentiment analysis methods account for the semantic orientation of individual words by matching words in a text with a sentiment lexicon, i.e., a list of words and their associated sentiment. The overall semantic orientation of a text is then determined by aggregating (e.g., summing) the sentiment scores of the individual words, as retrieved from the sentiment lexicon. In this sentiment scoring process, other aspects of content may be taken into account as well, such as negation (Heerschop et al., 2011*b*; Hogenboom et al., 2011), intensification (Taboada et al., 2011), or the rhetorical roles of text segments (Heerschop et al., 2011*a*; Hogenboom et al., 2010).

In machine learning approaches to sentiment analysis, text is typically represented as a vector. Such a vector denotes a bag-of-words, i.e., an unordered collection of words occurring in a document. Here, a binary encoding of text, indicating the presence of specific words, has proven to be an effective representation (Pang & Lee, 2004), outperforming a frequency-based vector representation of text (Pang et al., 2002). Vectors may also contain features other than words, e.g., parts of words, word groups, or features representing other aspects of content such as semantic distinctions between words (Whitelaw et al., 2005). Additionally, documents may be represented as a (small) bag-of-concepts (i.e., multi-word expressions of affective common-sense knowledge) (Cambria & Hussain, 2012) and features represented in vectors may be weighted as well (Paltoglou & Thelwall, 2010).

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., 2011*a*). Moreover, lexicon-based approaches can be generalised 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., 2011*a*) and are typically outperformed by machine learning approaches in terms of classification accuracy in specific domains for which machine learning approaches can be trained and optimised (Taboada et al., 2008).

This trade-off has inspired recent work to be focused on hybrid approaches, combining 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., 2012), 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. Only sentiment-carrying words are included in this vector representation, as these words are assumed to play a crucial role in conveying the overall sentiment of a text, 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.

2.2 Analysing Sentiment in a Multi-Lingual Setting

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 means of applying sentiment analysis techniques tailored to each specific considered language, as different sentiment analysis approaches are required for distinct languages (Moens & Boiy, 2007).

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 considered texts into a reference language, i.e., English. Subsequently, sentiment analysis is performed on the translated results. By doing so, Bautin et al. (2008) assume that the results of the analysis on both the original text and the translated text are comparable and that the errors made by the machine translation do not significantly influence the results of the sentiment analysis. However, the quality of the sentiment analysis on the translated text 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 clearly have their limitations, existing research on dealing with a multi-lingual setting when analysing sentiment typically deals with 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 (Hofman & Jijkoun, 2009; Wan, 2009) to constructing entirely new sentiment scoring frameworks (Abbasi et al., 2008; Dau et al., 2007*a*,*b*; Gliozzo & Strapparava, 2005; Moens & Boiy, 2007) 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 artefacts of that language rather than culture-free analytical tools (Wierzbicka, 1995*a*,*b*).

In this light, in order for language-specific sentiment scores to be meaningful when analysing 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 paper, 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 in natural language. Intuitively, both measures may be related to some extent, yet to the best of our knowledge, the relation between language-specific sentiment scores and universal sentiment classifications has not been previously investigated. The contribution of our current endeavours lies in investigating how language-specific sentiment scores can be mapped to universal star ratings.

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 conveyed sentiment. In addition, these 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., 2011*a*; Taboada et al., 2011). As such, these lexicon-based sentiment analysis techniques may prove to be useful for analysing 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 current endeavours.

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 analysing the sentiment of a document.

For each supported language, the sentiment analysis framework first prepares documents by cleaning the text – i.e., 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 negating or amplifying a sentiment-carrying word) from words that do not carry any sentiment.

Then, for each sentiment-carrying word t in a document d, the 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 t is not modified, m_t is set to 1.0. The 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 normalising the result for the number of sentiment-carrying words, i.e.,

$$\zeta_d = \frac{\sum_{t \in d} m_t \zeta_t}{|t \in d|}. \tag{1}$$

The normalised 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.

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 modelling the relation between language-specific 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 write rather positive texts. As such, texts belonging to, e.g., the four-star class should have higher sentiment scores than three-star texts. We thus map language-specific 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 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 1. One could expect one-star and five-star texts to be representing the extreme negative and positive cases, respectively, i.e., covering respective sentiment scores below -1 and above 1. The three-star class would intuitively be centred around a sentiment score of 0, indicating neutral or mixed sentiment. The two-star and four-star classes would then cover the remaining ranges of negative and positive scores, respectively, representing rather negative and positive conveyed sentiment, respectively.

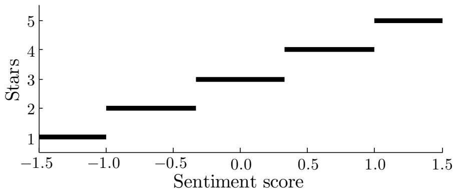
Many alternative mappings may exist for, e.g., different domains or languages. Mappings may for instance be skewed towards positive or negative sentiment scores or sentiment boundaries may be unequally spread across the full range of considered sentiment scores. In this light, the challenge is to find an optimal set of boundaries for each considered domain or language. The goal of this optimisation process is to minimise 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.,

$$\kappa_b = \sum_{i=1}^{5} \epsilon_{c_i}(b) . \tag{2}$$

As such, the optimisation 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.,

$$\zeta_l < b_1 < b_2 < b_3 < b_4 < \zeta_u . \tag{3}$$

Figure 1 Intuitive mapping from sentiment scores to universal star ratings.



Finding an optimal set of sentiment boundaries is not a trivial task, 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 may want to consider using a greedy algorithm when constructing a set of boundaries. Alternatively, heuristic or randomised optimisation techniques like genetic algorithms may be applied in order to explore the multitude of possible solutions. 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 analysed by means of sentiment analysis tools tailored to the language of these texts. The sentiment scores thus obtained can subsequently be transformed into universal 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.

3.3 Star Ratings as a Non-Monotonic Function of Sentiment Scores

The sentiment mapping as proposed in Section 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 modelling 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 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 modelling 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 naïve Bayes model and a support vector machine (Pang & Lee, 2008).

3.4 Incorporating a Bag-of-Sentiwords Representation

The machine learning methods for modelling sentiment mappings $M : \zeta_d \rightarrow c_d$ proposed in Section 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 naïve Bayes and support vector machine methods proposed in Section 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., 2012), 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. (2012), 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. (2012), we opt for a binary representation of our additional features Ξ, as we assume the sentiment conveyed by a text to be in the (number of) distinct words with a similar semantic orientation, rather than in their frequency of occurrence. Moreover, binary features have been shown to be more effective for sentiment analysis purposes than frequency-based features (Pang et al., 2002). In addition to incorporating a *bag-of-sentiwords* representation in our vector representations of sentiment-carrying text, 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., 2011*b*; Hogenboom et al., 2011).

As an example, 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 (possibly negated) words that convey this sentiment.

4 Evaluation

The methods proposed in Section 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.

4.1 Experimental Setup

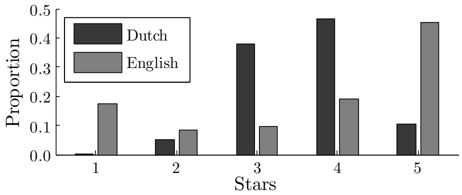
In our analysis, we consider two sets of similar documents. Our first data set consists of 1,759 short movie reviews in Dutch, crawled from various Web sites (Korte Reviews, 2011; Lemaire, 2011). The second data set considered in our current work consists of 46,315 short movie reviews in English (Metacritic, 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 data sets 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.

The documents in both data sets are first analysed for the sentiment conveyed by 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 analysing the sentiment of a document. For each supported language, the sentiment analysis framework performs the text cleaning, word tokenisation, POS tagging, word type classification, and sentiment scoring tasks as described in Section 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.

Using 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 yields 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 3.

Figure 2 Distribution of our Dutch and English documents over star rating classes.



First, we consider modelling the relation between language-specific 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 1. In order to optimise the location of the sentiment boundaries in these mappings, we use a brute force approach, where we optimise 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 modelling 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., 2012), i.e., a naïve 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) sentiment-carrying words in our Dutch documents and 884 features representing the occurrence of (negated) sentiment-carrying words in our English documents.

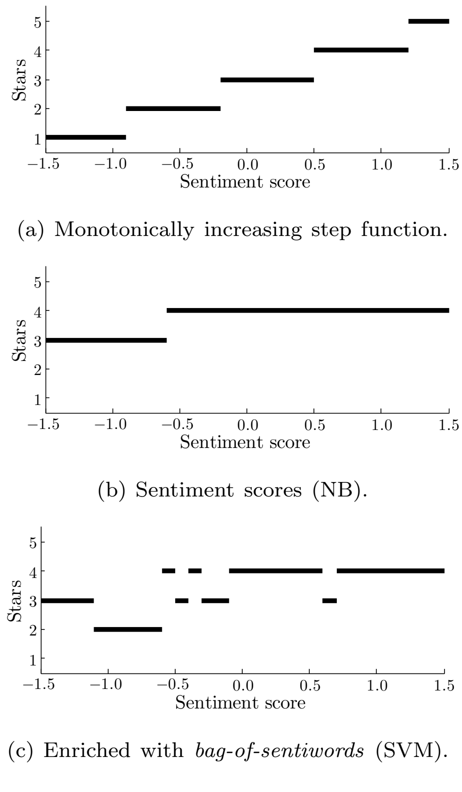
Our models' quality is assessed by means of the 10-fold cross-validated overall accuracy and the macro-level F_1-measure. Accuracy is the overall proportion of correctly classified documents. The macro-level F_1-measure is the average F_1-measure of the individual star rating classes, weighted for their 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, with probability distributions equal to the class distributions as depicted in Figure 2. We assess the statistical significance of performance differences by means of a paired two-sample one-tailed t-test.

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 visualised in Figure 3. Several observations can be made in these visualisations.

First, Figure 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 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.

Figure 3 Typical sentiment mappings constructed on the training set of one specific fold of our evaluation of our collection of opinionated Dutch documents. These sentiment mappings depict how the (majority of) our considered opinionated Dutch documents which convey specific sentiment scores in the test set of this fold 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).



(a) Monotonically increasing step function.

(b) Sentiment scores (NB).

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

Additionally, Figure 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 classes of intended sentiment 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 visualised in Figure 3(c). This suggests that the relation between conveyed and intended sentiment only partly 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 1, 2, and 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%.

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 collection of Dutch movie reviews, sentiment conveyed by a text as a whole can be better mapped to universal star ratings of intended sentiment when accounting for the specific sentiment-carrying words conveying the sentiment of the text. Nevertheless, even our best models can map conveyed sentiment to intended sentiment only to a limited extent, as our highest macro-level F_1-scores and overall accuracy are approximately 50%.

Table 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.

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 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	-0.400***	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 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 $p < 0.0001$.

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

4.3 Experimental Results on English Documents

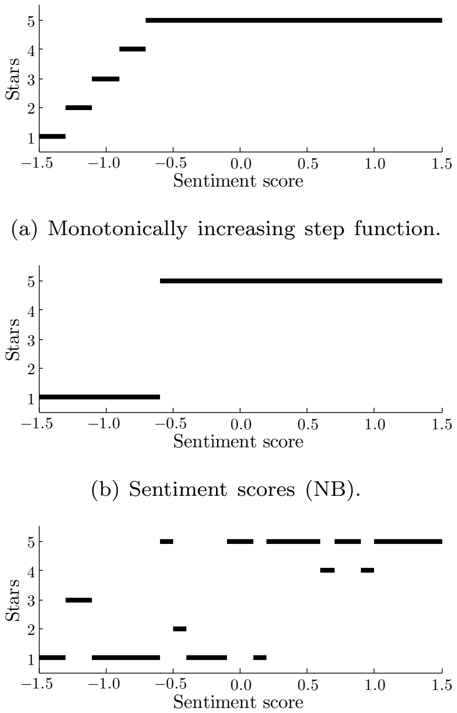
The sentiment mappings created using our methods on our collection of English movie reviews look different from the Dutch sentiment mappings. Figure 4 exhibits some of the distinctive features of our constructed sentiment mappings for our data set of English documents.

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 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.

Figure 4 Typical sentiment mappings constructed on the training set of one specific fold of our evaluation of our collection of English reviews. These sentiment mappings depict how the (majority of) our considered opinionated English documents which convey specific sentiment scores in the test set of this fold 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).



(a) Monotonically increasing step function.

(b) Sentiment scores (NB).

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

Tables 4, 5, and 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. 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.

Table 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.

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 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	-0.031***	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 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 $p < 0.0001$.

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

4.4 Overall Experimental Results

The results presented in Sections 4.2 and 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 rhetoric 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 reviews appears to be a better proxy for the intended sentiment. Another compelling example from our data is a review in which the author heavily cites and criticises 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 the (rhetorical) role of sentiment-carrying words in a text may in the future help improve our mappings of conveyed sentiment to intended sentiment.

5 Conclusions and Future Work

In this paper, we have proposed to use language-specific sentiment scores in order to classify natural language text into universal star ratings, capturing people's intended sentiment. We envisage such mappings to be useful in analytical tools for people's true sentiment, independent of domain, language, or culture. The results of our experiments with respect to modelling the relation between conveyed and intended sentiment for both a Dutch corpus and an English corpus suggest that the way natural language reveals people's intended sentiment may differ across distinct collections of documents. Additionally, the relation between conveyed and intended sentiment of documents in both considered data sets only partly depends on the sentiment conveyed by 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 paper indicate that, in practice, language-specific sentiment scores form a good starting point for capturing people's truly intended sentiment, when combined with the specific sentiment-carrying words constituting these scores. However, in order to be able to more accurately capture people's intended sentiment by means of analysing 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, in future work, we aim to 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, we aim to distinguish important from less important text segments with respect to the overall intended sentiment, for instance based on their rhetorical role, and to take this into account when classifying an author's intended sentiment. Additionally, the possibility for a user to provide star ratings when writing a review may give less incentive for a user to verbally express the sentiment associated with the rating. Therefore, we consider using other aspects of content, e.g., emoticons or latent cues, as a proxy for intended sentiment in future work.

As an alternative to our considered methods, future work could also be focused on using the occurrences of specific words, (latent) cues, and/or semantic and structural aspects of content in order to directly categorize text into universal classes of intended sentiment. Last, an interesting direction for future work could be to explore the cultural dimension of sentiment analysis, as people with different cultural backgrounds may make different use of the same language in order to convey their sentiment.

Acknowledgements

This article is a revised and expanded version of a paper entitled *Towards Cross-Language Sentiment Analysis through Universal Star Ratings*, presented at the Seventh International Conference on Knowledge Management in Organisations (KMO 2012), Salamanca, Spain, July 11–13, 2012. Special thanks go to

Teezir (http://www.teezir.com) for their data, technical support, and fruitful discussions. The authors of this paper are partially supported by the Dutch national program COMMIT.

References

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

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

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

Bautin, M., Vijayarenu, L. & Skiena, S. (2008), International Sentiment Analysis for News and Blogs, *in* '2nd International Conference on Weblogs and Social Media (ICWSM 2008)', AAAI Press, pp. 19–26.

Cambria, E. & Hussain, A. (2012), *Sentic Computing: Techniques, Tools, and Applications*, 2nd edn, Springer.

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

Cesarano, C., Dorr, B., Picariello, A., Reforgiato, D., Sagoff, A. & Subrahmanian, V. (2006), OASYS: An Opinion Analysis System, *in* 'AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs (CAAW 2006)', AAAI Press, pp. 21–26.

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

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

Dau, W., Xue, G., Yang, Q. & Yu, Y. (2007*a*), Co-Clustering Based Classification, *in* '13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)', Association for Computing Machinery, pp. 210–219.

Dau, W., Xue, G., Yang, Q. & Yu, Y. (2007*b*), Transferring Naive Bayes Classifiers for Text Classification, *in* '22nd Association for the Advancement of Articifial Intelligence Conference on Artificial Intelligence (AAAI 2007)', AAAI Press, pp. 540–545.

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

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

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

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

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

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

Heerschop, B., van Iterson, P., Hogenboom, A., Frasincar, F. & Kaymak, U. (2011*b*), Analyzing Sentiment in a Large Set of Web Data while Accounting for Negation, *in* 'Advances in Intelligent Web Mastering - 3, 7th Atlantic Web Intelligence Conference (AWIC 2011)', Vol. 86 of *Advances in Intelligent and Soft Computing*, Springer, pp. 195–205.

Hofman, K. & Jijkoun, V. (2009), Generating a Non-English Subjectivity Lexicon: Relations that Matter, *in* '12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)', Association for Computing Machinery, pp. 398–405.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F. & Kaymak, U. (2013), Exploiting Emoticons in Sentiment Analysis, *in* 'Twenty-Eighth Symposium on Applied Computing (SAC 2013)', ACM, pp. 703–710.

Hogenboom, A., Boon, F. & Frasincar, F. (2012), A Statistical Approach to Star Rating Classification of Sentiment, *in* 'Management Intelligent Systems, 1st International Symposium on Management Intelligent Systems (IS-MiS 2012)', Vol. 171 of *Advances in Intelligent Systems and Computing*, Springer, pp. 251–260.

Hogenboom, A., Hogenboom, F., Kaymak, U., Wouters, P. & de Jong, F. (2010), Mining Economic Sentiment using Argumentation Structures, *in* 'Advances in Conceptual Modeling - Applications and Challenges, 7th International Workshop on Web Information Systems Modeling (WISM 2010) at the 29th International Conference on Conceptual Modeling (ER 2010)', Vol. 6413 of *Lecture Notes in Computer Science*, Springer, pp. 200–209.

Hogenboom, A., van Iterson, P., Heerschop, B., Frasincar, F. & Kaymak, U. (2011), Determining Negation Scope and Strength in Sentiment Analysis, *in* '2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2011)', IEEE, pp. 2589–2594.

Jansen, B., Zhang, M., Sobel, K. & Chowdury, A. (2009), 'Twitter Power: Tweets as Electronic Word of Mouth', *Journal of the American Society for Information Science and Technology* **60**(11), 2169–2188.

Korte Reviews (2011), 'Korte Reviews'. Available online, `http://kortereviews.tumblr.com/`.

Lemaire (2011), 'Lemaire Film Reviews'. Available online, `http://www.lemairefilm.com/`.

Melville, P., Sindhwani, V. & Lawrence, R. (2009), Social Media Analytics: Channeling the Power of the Blogosphere for Marketing Insight, *in* '1st Workshop on Information in Networks (WIN 2009)'.

Metacritic (2011), 'Metacritic Reviews'. Available online, `http://www.metacritic.com/browse/movies/title/dvd/`.

Mihalcea, R., Banea, C. & Wiebe, J. (2007), Learning Multilingual Subjective Language via Cross-Lingual Projections, *in* '45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)', Association for Computational Linguistics, pp. 976–983.

Moens, M. & Boiy, E. (2007), 'A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts', *Information Retrieval* **12**(5), 526–558.

Paltoglou, G. & Thelwall, M. (2010), A study of Information Retrieval Weighting Schemes for Sentiment Analysis, *in* '48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)', Association for Computational Linguistics, pp. 1386–1395.

Pang, B. & Lee, L. (2004), A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts, *in* '42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)', Association for Computational Linguistics, pp. 271–280.

Pang, B. & Lee, L. (2008), 'Opinion Mining and Sentiment Analysis', *Foundations and Trends in Information Retrieval* **2**(1), 1–135.

Pang, B., Lee, L. & Vaithyanathan, S. (2002), Thumbs up? Sentiment Classification using Machine Learning Techniques, *in* 'Empirical Methods in

Natural Language Processing (EMNLP 2002)', Association for Computational Linguistics, pp. 79–86.

Short Reviews (2011), 'Short Reviews'. Available online, `http://shortreviews.net/browse/`.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011), 'Lexicon-Based Methods for Sentiment Analysis', *Computational Linguistics* **37**(2), 267–307.

Taboada, M., Voll, K. & Brooke, J. (2008), Extracting Sentiment as a Function of Discourse Structure and Topicality, Technical Report 20, Simon Fraser University. Available online, `http://www.cs.sfu.ca/research/publications/techreports/#2008`.

Van der Meer, J., Boon, F., Hogenboom, F., Frasincar, F. & Kaymak, U. (2011), 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', Association for Computing Machinery, pp. 765–772.

Wan, X. (2009), 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)', Association for Computational Linguistics, pp. 235–243.

Whitelaw, C., Garg, N. & Argamon, S. (2005), Using Appraisal Groups for Sentiment Analysis, *in* '14th ACM International Conference on Information and Knowledge Management (CIKM 2005)', Association for Computing Machinery, pp. 625–631.

Wiebe, J., Wilson, T., Bruce, R., Bell, M. & Martin, M. (2004), 'Learning Subjective Language', *Computational Linguistics* **30**(3), 277–308.

Wierzbicka, A. (1995*a*), *Alternative Linguistics: Descriptive and Theoretical Modes*, John Benjamins Publishing Company, chapter Dictionaries vs. Encyclopedias: How to Draw the Line, pp. 289–316.

Wierzbicka, A. (1995*b*), 'Emotion and Facial Expression: A Semantic Perspective', *Culture Psychology* **1**(2), 227–258.