Erasmus University Rotterdam

A Discourse Structure-Driven Approach for Sentiment Analysis of Text

Alexander Hogenboom

Full paper entitled “Polarity Analysis of Texts using Discourse Structure”, authored by Bas Heerschop, Frank Goossen, Alexander Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong, appeared in *Twentieth ACM Conference on Information and Knowledge Management (CIKM 2011)*, pages 1061–1070, ACM, 2011.

February 7, 2013

IS-SWIS 2013

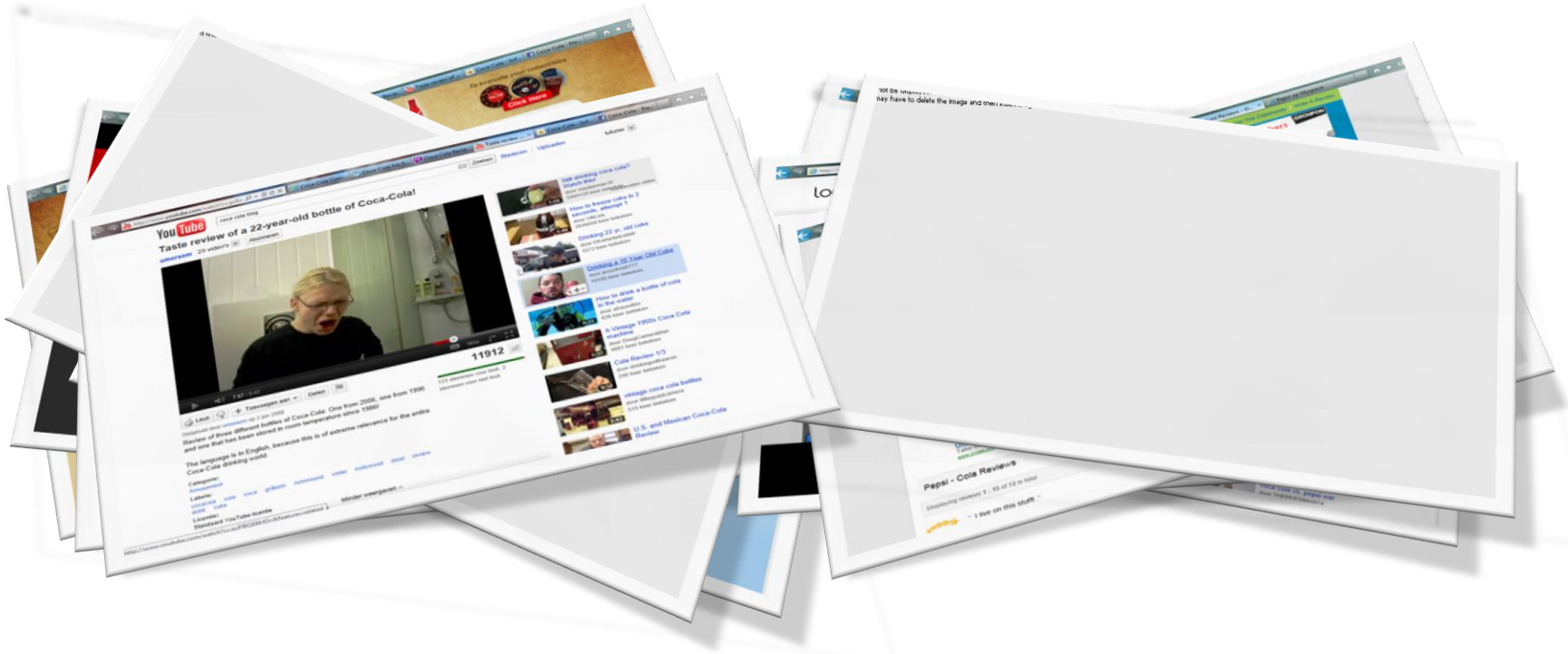
ERASMUS UNIVERSITEIT ROTTERDAM

Outline

- Introduction
- Sentiment Analysis
- Exploiting Structural Aspects of Text
- Framework
- Evaluation
- Conclusions
- Future Work

Introduction (1)

- Need for information monitoring tools for tracking sentiment in today's complex systems
- The Web offers an overwhelming amount of textual data, containing traces of sentiment



Introduction (2)

- An intuitive approach to sentiment analysis involves scanning a text for cues signaling its polarity
- Let us consider the following *negative* review:
  - Example: *Although Brad Pitt's well-deserved fall off a cliff was quite entertaining, this movie was terrible!*
- How can structural aspects of natural language text be exploited when determining its polarity?

Sentiment Analysis

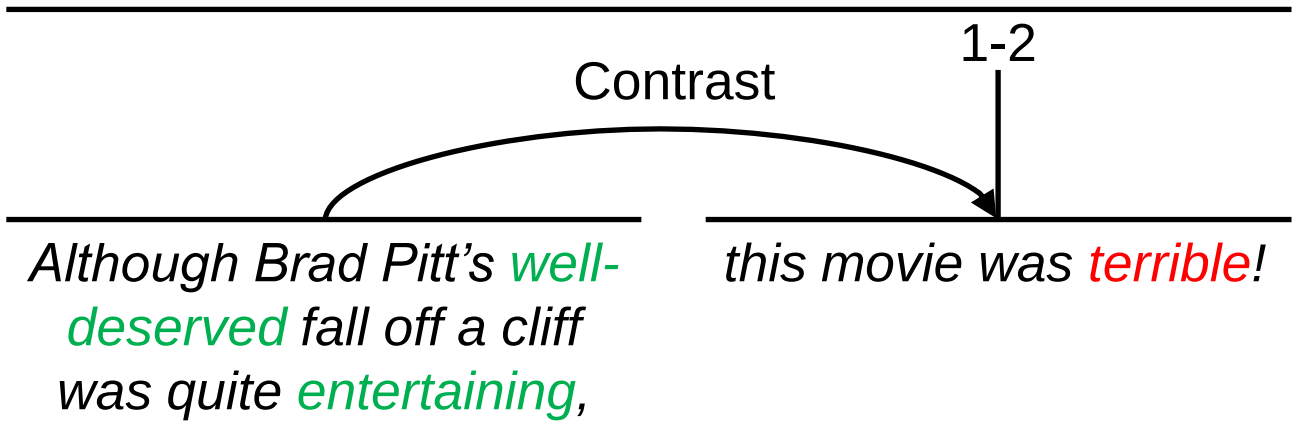
- Sentiment analysis is typically focused on determining the polarity of natural language text
- Applications in summarizing reviews, determining a general mood (consumer confidence, politics)
- State-of-the-art approaches classify polarity of natural language text by analyzing vector representations using, e.g., machine learning techniques
- Alternative approaches are lexicon-based, which renders them robust across domains and texts and enables linguistic analysis at a deeper level

Exploiting Structural Aspects of Text (1)

- Early approaches involve accounting for segments' positions in a text or their semantic cohesion
- Recent work exploits discursive relations by applying the Rhetorical Structure Theory (RST)
- RST can be used to split a text into a hierarchical structure of rhetorically related segments
- Nucleus segments form the core of a text, whereas satellites support the nuclei
- Many types of relations between segments exist, e.g., background, elaboration, explanation, contrast, etc.

Exploiting Structural Aspects of Text (2)

- Existing work differentiates between important and less important segments w.r.t. the overall sentiment
- Previously proposed method assigns different weights to nuclei and satellites
- We propose to differentiate among rhetorical roles, i.e., RST relation types



Framework (1)

- Lexicon-based document-level polarity classification
- Based on its lemma, Part-of-Speech (POS), and disambiguated word sense (Lesk-based), each individual word is scored in the range [-1,1] by using sentiment scores from SentiWordNet 3.0
- Word scores are aggregated and corrected for a bias towards positivity in order to classify text as positive (corrected score ≥ 0) or negative (corrected score < 0)
- Discourse parsing is applied in order to determine appropriate weights for word scores in this process

Framework (2)

- Simple discourse parsing: weights proportional to position of words in full text
- Sentence-level PArsing of DiscoursE (SPADE):
  - SPADE I:
    - Assign top-level RST nuclei weights of 1
    - Assign top-level RST satellites weights of 0
  - SPADE II:
    - Assign top-level RST nuclei weights of 1.5
    - Assign top-level RST satellites weights of 0.5
  - SPADE X:
    - Our novel, extended SPADE-based approach
    - Assign top-level RST nuclei and satellite types different weights in the range [-2, 2], optimized by means of a genetic algorithm

Evaluation (1)

- Implementation in Java, OpenNLP POS tagger, Java WordNet Library (JWNL) API for lemmatization, SentiWordNet 3.0 sentiment lexicon
- Corpus of 500 positive and 500 negative manually classified English movie reviews (60% training set, 40% test set)
- Baselines:
  - All words assigned weight of 1
  - Simple discourse parsing (position-based)
  - SPADE I
  - SPADE II

Evaluation (2)

- Alternative: SPADE X
- Optimized weight for nuclei equals 0.771
- Most significant RST relation weights:
  - Elaboration: 1.400
  - Enablement: 0.956
  - Attribution: 0.451
  - Contrast: -0.660

Evaluation (3)

	Positive			Negative			Overall	
Method	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Acc.**	**F1**
Baseline	.687	.690	.688	.688	.685	.687	.688	.687
Simple	.647	.660	.654	.653	.640	.647	.650	.650
SPADE I	.668	.695	.681	.682	.655	.668	.675	.675
SPADE II	.683	.690	.687	.687	.680	.683	.685	.685
SPADE X	**.732**	**.695**	**.713**	**.710**	**.745**	**.727**	**.720**	**.720**

Conclusions

- Recent sentiment analysis methods consider more and more aspects of content other than word frequencies
- We identify important and less important text segments and weight them accordingly when analyzing sentiment
- Both nuclei and satellites appear to play an important role in conveying sentiment, whereas satellites have until now been deemed predominantly irrelevant
- Significantly improved overall polarity classification accuracy and macro-level F1 w.r.t. not accounting for structural aspects of content comes at a cost of increased processing times

Future Work

- Account for full discourse (RST) trees rather than for top-level segmentations only
- Perform paragraph-level or document-level analysis of structure rather than sentence-level analysis
- Explore other (faster) methods of identifying discourse structure in natural language text
- Investigate our findings' applicability to vector-based machine learning approaches to sentiment analysis
- Evaluate our findings on different corpora

Questions?

Erasmus University Rotterdam

Alexander Hogenboom
Econometric Institute
Erasmus University Rotterdam
P.O. Box 1738, NL-3000 DR
Rotterdam, the Netherlands
hogenboom@ese.eur.nl