Rhetorical Structure Theory for Polarity Estimation: an Experimental Study

José M. Chenlo[a,*], Alexander Hogenboom[b], David E. Losada[a]

[a] *Centro de Investigación en Tecnoloxías da Información (CITIUS), University of Santiago de Compostela, Spain*
[b] *Econometric Institute, Erasmus University Rotterdam, The Netherlands*

Abstract

Sentiment Analysis tools often rely on counts of sentiment-carrying words, ignoring structural aspects of content. Natural Language Processing has been fruitfully exploited in Text Mining, but advanced discourse processing is still non pervasive for mining opinions. Some studies, however, extracted opinions based on the discursive role of text segments. The merits of such computationally intensive analyses have thus far been assessed in very specific, small-scale scenarios. In this paper, we investigate the usefulness of Rhetorical Structure Theory in various Sentiment Analysis tasks on different types of information sources. First, we demonstrate how to perform a large-scale ranking of individual blog posts in terms of their overall polarity, by exploiting the rhetorical structure of a few key evaluative sentences. In order to further validate our findings, we additionally explore the potential of Rhetorical Structure Theory in sentence-level polarity classification of news and product reviews. Our most valuable polarity classification features turn out to capture the way in which polar terms are used, rather than the sentiment-carrying words per se.

Keywords: Information Retrieval, Text Mining, Sentiment Analysis, Natural Language Processing, Rhetorical Structure Theory

1. Introduction

Natural Language Processing (NLP) has become of vital importance for current information systems [1]. Recent advances in NLP permit to distill actionable

*Corresponding author. Tel.: +34 881 816 396; Fax: +34 881 816 405.
Email addresses: josemanuel.gonzalez@usc.es (José M. Chenlo), hogenboom@ese.eur.nl (Alexander Hogenboom), david.losada@usc.es (David E. Losada)

knowledge from massive amounts of Web content: by detecting important events in news [2]; by discovering topics and viewpoints from social media [3, 4, 5]; or by associating news messages and social media posts with their potential effects on stock prices [6, 7] or sales [8].

Sentiment Analysis (SA) –also known as Opinion Mining– is an active and influential research area concerned with automatically extracting subjectivity from natural language text [9, 10, 11, 12]. It deals with tasks such as classifying the polarity of documents as positive or negative, or ranking documents in terms of their associated degree of positivity or negativity with respect to a topic of interest. With one fifth of all tweets [13], one third of all blog posts [14], and the vast majority of reviews discussing products or brands, SA is crucial for revealing traces of people's sentiment from ubiquitous user-generated content.

Commercial sentiment analysis systems mostly rely on simple occurrences of sentiment-carrying words in texts [12]. Yet, word frequencies alone are insufficient for mining sentiments [12, 15, 16, 17]. Accounting for the way in which words are used is essential for text understanding. In this light, one of the key open research issues refers to the role of textual structure [12]. Structural aspects seem to be valuable for various Text Mining tasks [18, 19, 20, 21], including SA [15, 16, 22, 23, 24], but this requires further study.

An increasingly popular way of accounting for structural aspects of opinionated text is to analyse the rhetorical organization of documents [16, 24, 25, 26, 27, 28]. One way of accomplishing this is by applying the Rhetorical Structure Theory (RST) [29]. As a leading descriptive framework for text, RST identifies the rhetorical roles (e.g., explanation, contrast) of text segments. This is useful for SA, as sentiment-carrying words in, e.g., an explanation segment may contribute differently to the overall sentiment than those in a contrasting segment. But RST for SA has been mostly evaluated in small-scale document classification studies [16, 24]. In such constrained settings, RST significantly contributed to determine polarity, at a cost of computational complexity.

In this paper, we thoroughly study RST for SA and experiment with various opinion repositories –which vary in size and source, i.e., blogs, news, and product reviews– and different polarity analysis tasks of varying granularity, i.e., document polarity ranking and sentence-level polarity classification. To the best of our knowledge, this is the first study of this kind.

In the first part of this paper, we quantify the advantage of exploiting RST for blog post polarity ranking and we identify the rhetorical relations that help to understand the sentiment conveyed in blogs. For efficiency reasons, we build upon recent advances in extracting key opinionated sentences from blog posts [30] and analyse the discourse only for selected passages. The evaluation of RST is therefore indirect, as we test how the rhetorical analysis of selected sentences helps

to estimate the polarity of documents as a whole.

In the second part of this paper we focus on a fine-grained task –sentence-level polarity classification– and perform a more direct evaluation of the merits of RST. We study the ability of RST to reveal positively or negatively-oriented sentences within news articles or product reviews. Sentence-level polarity classification goes beyond document-level sentiment classification as it moves closer to opinion targets and sentiments on the targets. This facilitates opinion extraction from text that may only contain a few sentences that discuss the topic of interest [10].

The remainder of this paper is organised as follows. First, in Section 2, we discuss related work on existing SA approaches and review how structural aspects of content are typically involved in such methods. In Section 3, we describe our novel method of document-level polarity estimation that works at large scale. Section 4 focuses on sentence-level polarity classification guided by RST. Last, in Section 5, we present our conclusions and suggest directions for future research.

2. Related Work

Explicit information on user opinions is often hard to find, confusing, or overwhelming [9]. The abundance of user-generated content on the Web has led to a surge of research interest in systems that automatically mine opinions and sentiment. Many of such SA systems exist, but the exploration of how to account for structural aspects of content when analysing sentiment has only just begun.

2.1. Sentiment Analysis

SA tools often apply Computational Linguistics and Text Mining technology. Typical tasks include distinguishing subjective text segments from objective ones, as well as determining the polarity of words, sentences, or documents [9]. The latter is often approached as a two-class categorisation problem of distinguishing positive from negative text or, occasionally, as a three-class problem, in which an additional class of neutral text is considered. An alternative to polarity classification is to determine a degree of positivity or negativity of natural language text and produce, e.g., rankings of positively and negatively oriented documents.

The state-of-the-art in automated SA has been reviewed extensively [9, 10, 11, 12]. Existing methods range from Machine Learning methods, exploiting patterns in vectorial representations of text, to lexicon-based methods, accounting for the semantic orientation of individual words (e.g., using sentiment lexicons). Many hybrid approaches exist as well.

Large-scale SA tasks typically pose unique challenges, not just in extracting sentiment from a large set of documents, but also in identifying on-topic (fragments of) documents. Numerous studies have been conducted on how to mine opinions

from large-scale repositories like the blogosphere. Given a topic of interest, the search for relevant and subjective documents (regardless of their polarity) has been studied by different scientists [31, 32, 33]; and Chenlo and Losada proposed effective and efficient methods of finding opinionated segments in blog posts [30]. These methods permit to represent the overall opinion of a blog post with a limited number of sentences that are selected by combining three types of sentence-level evidence: topicality, polarity, and location.

Although relying on counts of sentiment-carrying words is still predominant [12], other aspects of content are promising. Early work with movie reviews [22, 34] considered the absolute position of text segments and found that the last sentences of a document could be indicative of the overall polarity. Positional information has proven to be useful in large-scale SA tasks as well. For example, the proximity of query terms to subjective sentences in a document was used by Santos et al. to detect on-topic opinions [32]. Similarly, Gerani et al. defined a proximity-based propagation method to calculate the aggregated opinion at the position of each query term in a document [31].

A broad array of studies employed linguistic mechanisms to extract structure. For instance, Devitt and Ahmad estimated sentiment by analysing the semantic cohesion of a text [23], with limited success. More successful attempts [16, 24, 26, 27, 28] selected important opinion extracts from the text's rhetorical structure –obtained by, e.g., RST.

2.2. Rhetorical Structure Theory

The structure of natural language can be characterised by the rhetorical relations that hold between parts of the text. Such relations (e.g., explanations or contrasts) are important for text understanding, because they give information about how the textual segments are related to one another to form a coherent discourse. 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.

Rhetorical Structure Theory [29] is one of the leading discourse theories. The theory can be used to split text into rhetorically related segments. Each segment may in turn be split as well, thus yielding a hierarchical 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; the main paper on RST defines 23 types of relations [29]. 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.

For example, the core of the sentence "*Although I like the characters, the book is horrible*" provides a negative sentiment with respect to a book ("*the book is horrible*"). The other segment is a satellite with contrasting information with respect to the nucleus, admitting to some positive aspects of the book. For a human reader, the polarity of this sentence is clearly negative. However, in a classical (word-counting) approach, all words would contribute equally to the total sentiment, thus yielding a verdict of a neutral or mixed polarity at best. Accounting for the rhetorical structure could result in shifting the focus of the analysis to the nucleus segment, e.g., by giving the nucleus a higher weight. Furthermore, distinct rhetorical relations could be treated differently in the SA process.

2.3. Rhetorical Structure in Sentiment Analysis

In automated SA, rhetorical relations are typically exploited to distinguish important text segments from less important ones in terms of their contribution to a text's overall sentiment. Early work made crude distinctions between nuclei and satellites, with satellites being assigned relatively low weights, or no weight at all [24]. More recently, Heerschop and colleagues successfully differentiated between distinct types of satellites when assigning weights to text segments [16]. Discourse relations have also proven valuable to eliminate polarity ambiguities or to aggregate sentiments from neighbouring segments. Zhou et al. [25] designed a rule-based method that disambiguates the sentiment of text segments containing conflicting opinions. Zirn et al. [26] adopted Markov logic to integrate polarity scores –obtained from various lexicons– with neighbourhood and discursive information.

The papers discussed above rely strongly on rhetorical relations, as identified by applying RST. Other studies in the literature [27, 28] rely on other types of discursive cues for SA. Taken as a whole, the body of work on discourse analysis for SA demonstrates the viability of using rhetorical structure. However, one of the reported downsides of SA guided by rhetorical structure is the high processing time required for analysing discourse in natural language text [16]. This problem seems to thwart the applicability of such methods in large-scale scenarios. Yet, this problem can be mitigated by combining RST and search. In this paper, we follow this path and apply RST on a limited number of sentences selected from blog posts. The sentences are extracted by employing effective and efficient methods to determine opinions within blog posts [30].

3. Sentiment-Based Ranking of Blog Posts using Rhetorical Structure Theory

Several efforts have recently been conducted to detect opinions in blog documents [35]. Search alone is not sufficient for building information systems that effectively deal with the opinionated nature of data in the blogosphere [9]. Large-scale opinion mining is typically considered as a two-stage process that involves an initial topic retrieval stage for retrieving relevant documents given a user query, and a subsequent re-ranking stage accounts for opinion-based features [36]. 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 documents related to the query, and a subsequent polarity estimation task that aims to identify the orientation of a document with respect to the topic (e.g., positive or negative).

The latter polarity estimation task poses many unresolved issues, such as dealing with irony or conflicting opinions. Yet, typical approaches naively estimate the polarity of a text based on the frequencies of positive and negative terms [12]. In line with recent findings [12, 15, 16, 17], we argue that the polarity estimation problem cannot be solved with regular matching (or count-based) techniques alone. In fact, most lexicon-based methods have failed to retrieve more positive or negative documents than baselines that do not account for the polarity of individual terms at all [30, 36].

Rhetorical roles of text segments and their relative importance should be accounted for when determining the overall polarity of a text [16, 24, 26, 27, 28]. But such rhetorical analysis is computationally demanding and applying it at large scale is challenging. We have addressed this challenge for a polarity ranking task and, in this section, we report on the rhetorical relations that give good guidance for understanding the sentiment conveyed by blog posts. We also quantify the advantage of exploiting these relations and compare our RST-based methods with conventional approaches for large-scale polarity ranking of blog posts.

3.1. Method

In the blogosphere, the presence of spam, off-topic information, or relevant yet non-opinionated content is a major issue that harms the effectiveness of opinion finding techniques. Therefore, we propose to build upon recent advances in extracting key opinionated sentences for polarity estimation in blog posts [30], by concentrating on the structure of the discourse of those selected sentences. This is beneficial to avoid noisy chunks of text and also convenient from a computational complexity perspective.

We first present the method for finding polar sentences in blog posts (Section 3.1.1). Then, in Section 3.1.2, we elaborate on how to perform rhetorical analysis over these key evaluative sentences. Last, in Section 3.1.3, we explain how to

determine the overall orientation of individual blog posts by analysing these key evaluative sentences and their associated rhetorical structure.

3.1.1. Finding Relevant Polar Sentences

Given a ranked list of documents (ordered by decreasing topic relevance, $\text{rel}(D, Q)$, with respect to a query Q), we search for on-topic opinions as follows. The polarity $\text{pol}(S, Q)$ of a sentence S with respect to Q is defined as a linear combination of the normalised relevance of the sentence and the polarity of the sentence:

$$\text{pol}(S, Q) = \beta \cdot \text{rel}_{\text{norm}}(S, Q) + (1 - \beta) \cdot \text{pol}(S), \tag{1}$$

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

The normalised relevance of a sentence S with respect to a query Q is computed as a normalised tf-idf score:

$$\text{rel}_{\text{norm}}(S, Q) = \frac{\text{tf-idf}(S, Q)}{\max_{S_C} \text{tf-idf}(S_C, Q)}, \tag{2}$$

where the maximum $\max_{S_C} \text{tf-idf}(S_C, Q)$ is computed over all sentences S_C from the ranked list of documents for the query Q.

We use the Lemur[1] implementation of tf-idf, with BM25-like weights, for sentence retrieval:

$$\text{tf-idf}(S, Q) = \sum_{t \in Q \cap S} \frac{k_1 \cdot f_{t,S}}{f_{t,S} + k_1 \cdot \left((1 - b) + b \cdot \frac{|S|}{l_c}\right)} \cdot \frac{N}{n_t} \cdot f_{t,Q}, \tag{3}$$

where $f_{t,S}$ is the frequency of the term t in the sentence S; $f_{t,Q}$ signals the frequency of the term t in the query Q; $|S|$ and l_c capture the length of the sentence and the average length of sentences in the collection, respectively (computed as the number of words); N is the number of sentences in the collection; n_t is the number of sentences that contain the term t; and k_1 and b are free parameters. This is a sentence-query scoring function that defines both how matching terms are weighted and how the weights are combined to give a sentence score. It incorporates typical Information Retrieval elements, such as term frequency, inverse sentence frequency and sentence length normalisation. The term frequency component grows with the number of ocurrences of the term in the sentence ($f_{t,S}$). This is based on the assumption that high frequency terms are important for describing the key topics of a text. BM25 models term frequency as a function that starts at

[1] `http://www.lemurproject.org/`

zero rises steeply at first, and then flattens out to reach an asymptotic limit. The parameter k_1 controls the speed of the approach to this limit. The factor N/n_t represents an inverse sentence frequency weight. This captures term specificity in the sentence collection. Matching terms that occur in few sentences are preferred. The combined effect of term frequency and inverse sentence frequently makes that the high scoring terms are those that are repeated within a narrow chunk of text but infrequent in the collection as a whole. Length normalisation is modelled by $\left((1-b)+b\cdot\frac{|S|}{l_c}\right)$. This tries to avoid long sentences having unfair advantage in the retrieval process. Long sentences have more unique terms and higher average term frequency (more repetitions). If we do not account for length normalisation we would be unfairly penalising shorter chunks of text. A b value of 1 means that the weight is fully normalised by the sentence length ($|S|$), whereas $b = 0$ means no normalisation. In our experiments, we applied the well-known BM25 suggested configuration, i.e., $k_1 = 1.2$ and $b = 0.75$ [37].

The second component of the polarity of a sentence S with respect to a query Q, as defined in equation 1, is the polarity $\text{pol}(S)$ of the sentence S. We propose to compute this score based on OpinionFinder (OF) [38], which is a well-known system for subjectivity analysis. OF estimates which sentences in a document are subjective and, additionally, marks various aspects of subjectivity in those sentences. These aspects include the source (i.e., holder) of the opinions, as well as the words that are included in phrases expressing positive or negative sentiments. The output produced by OF has been shown to be useful in numerous subjectivity and polarity studies [30, 32, 33, 39].

Following existing work [30], we define $\text{pol}(S)$ as the number of positive (or negative) terms tagged by OF, divided by the total number of terms in a sentence S. When retrieving positive documents, $\text{pol}(S)$ captures the proportion of positive terms in the sentence, whereas in case of a negative document retrieval scenario, $\text{pol}(S)$ quantifies the proportion of negative terms in the sentence.

The polarity of a document D can be computed based on the polarity scores $\text{pol}(S, Q)$ of some sentences selected from D. The aggregated score, $pol_S(D, Q)$, can be defined as the average score of *all* polar sentences, the average score of the *first* or *last* k polar sentences, or the average score of the k sentences with the *highest* polarity score [30]. The latter method, referred to as *PolMeanBestN*, is robust and outperforms the alternatives [30]. Therefore, we adopt *PolMeanBestN* to identify the key evaluative sentences. Since $k = 1$ was shown to be the best configuration [30], we extract a single polar sentence –i.e., the one with the highest $\text{pol}(S, Q)$– to estimate the overall polarity of the blog post.

Relation	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 1: Rhetorical Structure Theory relations taken into account in our study.

3.1.2. Sentence-Level Parsing of Discourse

The sentence-level polarity scores, $\mathrm{pol}(S)$, may be inaccurate. Besides word frequencies, the sentence's discourse units and their interrelations are meaningful for sentiment estimation.

Given the key evaluative sentences, we use a tool for Sentence-level PArsing of DiscoursE (SPADE) [40], which creates RST trees for individual sentences. SPADE was trained and tested on the train and test set of the RST Discourse Treebank (RST-DT) [41], achieving an F_1 score of 83.1% on identifying the right rhetorical relations and their right arguments [40]. The discourse relations that we take into account are detailed in Table 1.

To exploit the discourse structure of the key evaluative sentences we assign distinct polarity weights to specific parts of the sentence. This is guided by the identified rhetorical roles of the sentence's parts. Following existing work [16, 24], 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 redefine the polarity score $\mathrm{pol}(S)$ of a key evaluative sentence S as a weighted sum of the polar terms occurring in the top-level nucleus and satellite, respectively, i.e.,

$$\mathrm{pol}(S) = w_{\mathrm{nuc}} \cdot \mathrm{pol}_{\mathrm{nuc}}(S) + w_{\mathrm{sat}} \cdot \mathrm{pol}_{\mathrm{sat}}(S), \tag{4}$$

where w_{nuc} is the weight for the top-level nucleus, w_{sat} is the weight for the top-level satellite; and $\text{pol}_{\text{nuc}}(S)$ and $\text{pol}_{\text{sat}}(S)$ represent the proportion of positive or negative terms tagged in the nucleus and satellite, respectively.

Observe that w_{nuc} and w_{sat} are free parameters that need to be trained for each distinct rhetorical relation. For instance, the sentence *"the film was awful but it was nice going there with her"* could get a satellite weight that is smaller than the nucleus weight. The weight of the satellite (*"but it was nice going there with her"*) could even be negative here in order to account for the contrasting relationship between the two segments. Conversely, in the sentence *"the film was awful, and especially the acting was downright horrible!"*, the satellite (*"and especially the acting was downright horrible!"*) elaborates on the matters expressed in the nucleus and should hence receive a different treatment.

Combined with the relevance of each key evaluative sentence S with respect to a query Q, the polarity score $\text{pol}(S)$ thus computed provides an intricate estimation of the polarity $\text{pol}(S, Q)$ of S with respect to Q (equation 1).

3.1.3. Blog Post Ranking Guided by the Rhetorical Structure of Key Sentences

These sentence-level polarity scores can subsequently be used to estimate the polarity $\text{pol}(D, Q)$ of their associated document D with respect to the query Q:

$$\text{pol}(D, Q) = \gamma \cdot \text{rel}_{\text{norm}}(D, Q) + (1 - \gamma) \cdot \text{pol}_S(D, Q), \tag{5}$$

where $\gamma \in [0, 1]$ is a free parameter, and $\text{rel}_{\text{norm}}(D, Q)$ is the relevance score of document D with respect to query Q after a query-based normalisation in the interval $[0, 1]$. Under *PolMeanBestN* with $k = 1$, the sentence-based polarity $\text{pol}_S(D, Q)$ is defined as:

$$\text{pol}_S(D, Q) = \max_{S \in D} \text{pol}(S, Q), \tag{6}$$

Documents are re-ranked by decreasing $\text{pol}(D, Q)$. This promotes on-topic blog posts that are positively or negatively opinionated. We did not optimise the parameters β (equation 1) and γ (equation 5). We simply set them to the configuration suggested in [30]: $\beta = 0.6$ and $\gamma = 0.6$ for negative polarity estimation; and $\beta = 0.2$ and $\gamma = 0.5$ for positive polarity estimation.

3.2. Experiments

We have evaluated our RST-based polarity ranking method on a large-scale multi-topic dataset, i.e., the BLOGS06 text collection [42]. This corpus is one of the most renowned blog test collections with relevance, subjectivity, and polarity assessments.

3.2.1. Collection and Topics

We employed the benchmarks of the TREC 2006, TREC 2007, and TREC 2008 blog tracks. All of these tracks had the BLOGS06 corpus as reference collection (3,215,171 documents). Each year, a new set of 50 topics was provided and new judgments were made according to the documents retrieved by the participants. Documents were judged by TREC assessors in two different aspects. First, documents were judged on topic relevance –a post can 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.

In the polarity ranking task, systems had to return a ranking of positive blog posts and a ranking of negative blog posts (related to a given query). Each query topic contained three different fields, i.e., title, description, and narrative. We only considered the title field, which is short and representative of real users' Web queries [36], and preprocessed documents and topics by stemming with the Krovetz stemmer and by removing 733 English stopwords.

3.2.2. Retrieval Task and Polarity Baselines

In TREC 2008, to allow the study of the performance of a specific opinion-finding technique across a range of different topic-relevance baseline systems, a set of five topic-relevance baselines was provided. These standard baselines use a variety of different retrieval approaches, and have varying retrieval effectiveness. The baselines were selected by TREC from the runs submitted to the initial ad-hoc retrieval task in the TREC blog track.

Spam detection, topic retrieval in blogs, and subjectivity classification are out of the scope of this paper. We focused on the polarity ranking task and applied our proposed RST-guided method to the set of subjective documents identified by the standard baseline runs. The input to our methods is therefore a set of opinionated documents with varied polarity orientations (positive, negative, or mixed polarity); and the objective is to search for positive documents and for negative documents. This polarity task per se is quite challenging because blog posts contain many offtopic passages and conflicting opinions. We evaluated performance in terms of Mean Average Precision (MAP) and Precision at 10 documents (P@10). These measures are commonly applied in IR to evaluate effectiveness of ranked outputs. We estimated statistical significance with the paired t-test at the 95% significance level.

3.2.3. Training and Test

We built a realistic and chronologically organised query dataset with the TREC topics: the TREC 2006 and TREC 2007 topics were used for training, and the TREC 2008 topics were used for testing. Two different training-test processes were run: one for positive polarity retrieval and another for negative polarity retrieval. The performance measure optimised was MAP.

On our training set, we optimised the weights of the rhetorical elements, used in equation 4. The weight of nuclei was fixed to 1 in order to reflect the alleged importance of these elements, whereas weights of satellites were assumed to be real numbers in the interval $[-2, 2]$. In this way, we allowed for satellites to contribute positively or negatively to the overall sentiment, as well as to be more important or less important than nuclei. For parameter tuning we employed Particle Swarm Optimisation (PSO), which is an effective method for automatically tuning parameters [43]. More details on this method can be found in Appendix A.

3.2.4. Experimental Results

Table 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 SA. Note that this polarity task is quite challenging –most TREC polarity systems failed to retrieve more positive or negative documents than the baselines [36]. Our results exhibit several patterns:

Polarity retrieval performance. *PolMeanBestN* estimates the overall recommendation of a blog post by taking into account the on-topic sentence in the blog post that has the highest polarity score. Previous studies reported *PolMeanBestN* to be comparable to the best performing approach at the TREC 2008 Blog track (KLE system) [30, 35]. Our novel method that combines *PolMeanBestN* with RST-guided SA is the best performing approach in our current experiments, typically showing significant improvements with respect to both the baseline and *PolMeanBestN*.

Positive versus negative retrieval. The effectiveness of all methods on negative document rankings is lower than on positive document rankings. This may be caused by negative documents being harder to find in the blog post collection. There are more positive documents than negative documents in the polarity judgments, i.e., 3,338 positive documents against 2,789 negative ones. Additionally, the lexicon-based identification of negative documents may be thwarted by people having a tendency of using rather positive words to express negative opinions [16].

Weights for rhetorical relations. Table 3 reports the weights learned for distinct RST elements. Having been assigned a weight of 1, nuclei are assumed to play a more or less important role in conveying the overall sentiment of a piece of

	Positive		Negative	
Method	MAP	P@10	MAP	P@10
Baseline1	.266	.368	.240	.296
+*PolMeanBestN*	.270	.372	.241	.300
+*PolMeanBestN(RST)*	**.273**	**.374**△ ▲	**.252**	**.318**△ ▲
Baseline2	.239	.334	.217	.278
+*PolMeanBestN*	.236	.316	.222	.282
+*PolMeanBestN(RST)*	**.242**△	**.356**△ ▲	**.226**▲	**.310**△ ▲
Baseline3	.276	**.350**	.249	**.284**
+*PolMeanBestN*	.276	.342	.252	.276
+*PolMeanBestN(RST)*	**.277**△	.338▼	**.258**△ ▲	.282
Baseline4	**.273**	.358	.264	.274
+*PolMeanBestN*	.271	.350	.273	.284
+*PolMeanBestN(RST)*	.272	**.362**△ ▲	**.283**△	**.324**△ ▲
Baseline5	.239	.360	.224	.300
+*PolMeanBestN*	.240	.358	.228	.312
+*PolMeanBestN(RST)*	**.279**△ ▲	**.438**△ ▲	**.239**	**.342**△ ▲

Table 2: Polarity ranking results. Mean average precision (MAP) and Precision at 10 (P@10) for positive and negative rankings of blog posts. The symbols ▲ (▼) and △ (▽) indicate significant improvements (decreases) over the original baselines provided by TREC and the *PolMeanBestN* method, respectively. The best value in each column for each baseline is bolded.

natural language text. Yet, some types of satellites appear to play an important role as well in conveying a text's overall sentiment. For instance, the most salient relations (i.e., those having the highest frequency of occurrence) in our training set are the *elaboration* and the *attribution* relations. For both positive and negative documents, satellite segments elaborating on matters presented in nuclei are typically assigned relatively high weights, exceeding those assigned to nuclei (w_{sat} = 2.00). Bloggers may, therefore, tend to express their sentiment in a more apparent fashion in elaborating segments rather than in the core of the text itself. A similar pattern emerges for attributing satellites as well as for persuasive text segments, i.e., those involved in *enablement* relations, albeit to a more limited extent (lower frequency of occurrence). Interestingly, however, the information in attributing satellites appears to be more important in negative documents than in positive documents. Another important observation is that the sentiment conveyed by elements in contrast satellites gets a negative weight. This permits to reverse the polarity score of contrasting segments, e.g., of the second part of the sentence *"The film was awful but it was nice going with her"*.

Relation	Positive Occurrence (%)	w_{sat}	Negative Occurrence (%)	w_{sat}
Attribution	18.3	0.531	17.7	2.000
Background	3.4	-0.219	3.8	-2.000
Cause	0.9	1.218	0.9	-0.011
Comparison	0.3	-1.219	0.3	-2.000
Condition	2.9	-0.886	2.5	-2.000
Consequence	0.1	0.846	0.1	1.530
Contrast	1.6	-1.232	1.7	-2.000
Elaboration	20.7	2.000	21.9	2.000
Enablement	3.8	2.000	3.8	1.221
Evaluation	0.1	0.939	0.1	-2.000
Explanation	0.7	2.000	0.8	2.000
Joint	0.9	-1.583	1.0	1.880
Otherwise	0.1	-1.494	0.1	-0.428
Temporal	0.3	-2.000	0.3	-0.448

Table 3: Satellite weights (w_{sat}) for RST relation types trained with PSO over positive and negative rankings, along with the occurrence rate of the relation types in the training data. The weight of the nuclei for all RST relations (w_{nuc}) was set to 1. The most salient relations (highest % of occurrence) are printed in bold.

4. Sentence Polarity Classification using Rhetorical Structure Theory

In the previous section, we described how to exploit the rhetorical structure of some selected sentences in order to estimate the overall polarity of blog posts. Effectiveness was not only influenced by RST, but also by the methods applied to extract opinionated sentences. As such, this gives us only indirect evidence of how RST helps to understand the polarity of natural language text. In this section we describe new experiments performed to evaluate RST in a more direct way.

We explored the potential of RST for a fine-grained task: sentence polarity classification. Since the TREC blog collection lacks sentence-level polarity judgements, we considered document repositories from other domains, i.e., news and product reviews. To this end, we selected the English documents from the Multilingual Opinion Analysis Task (MOAT) [44] and the Finegrained Sentiment Dataset (FSD) [45], both of which contain relevance and opinion judgements at sentence level.

4.1. Method

Sentence polarity classification is concerned with assigning a polarity label, e.g., positive or negative, to sentences. This categorisation task can be done by an automatic classifier that is constructed from training data (sentences represented

as feature vectors and their polarity labels). The characteristics of sentences can be very well encoded as features in a vector representation that is the input of the classifier. In our experiments, we considered the following sets of features:

- ***Vocabulary features.*** These are binary features based on the occurrence of unigrams and bigrams in the sentence. We only represented unigrams and bigrams that occur at least four times in the collection. Unigrams and bigrams are often valuable to detect specific domain-dependent (opinionated) expressions. The discriminative power of this type of content features has been shown in several opinion mining studies [22, 46].

- ***Length features.*** These features encode the length (number of words) of the sentence, the top-level nucleus, and the top-level satellite. The length of these text spans could be indicative of subjectivity or objectivity (e.g., factual sentences may be shorter). We also included the length of the document the sentence originates from as an additional sentence feature, as shorter documents –e.g., press releases– may be more factual than longer ones.

- ***Positional features.*** Positional evidence might benefit the polarity classification process as well. Recent results indicated that the position of sentences in a document is an important cue for the overall polarity [22, 34, 30]. We included features that encode the absolute position of the sentence within the document (e.g., 2 for the second sentence in the document), and the relative positions (i.e., absolute positions normalised by the number of sentences in the document).

- ***RST features.*** The presence of specific types of satellites may serve as a valuable cue for polarity. As we discussed in Section 3, the rhetorical roles of text segments can effectively guide the opinion detection process. For example, an attribution relationship could be indicative of subjectivity. Therefore, we included one binary feature for each type of RST relationships (Table 1). Every sentence has only one of these features set to 1 (determined by the sentence's top-level nucleus-satellite relationship).

- ***Sentiment RST features.*** These features are based on counting the positive and negative terms that occur in the nucleus and in every type of satellite. In this way, we individually represent the positivity and negativity of the nucleus, an attribution satellite, a contrast satellite, and so forth. Again, the representation is sparse because every sentence only contains one (top-level) satellite type. The positive and negative terms were obtained from the OF [38] sentiment lexicon. We included absolute and relative counts

Set	Feature
Vocabulary	Unigrams and bigrams (binary)
Length	Length of the sentence
	Length of the nucleus
	Length of the satellite
	Length of the document that contains the sentence
Positional	Absolute position of the sentence in the document
	Relative position of the sentence in the document
RST	Contains a satellite (binary)
	Contains specific satellite types (binary)
Sentiment RST	Number of positive terms in the nucleus
	Number of negative terms in the nucleus
	Number of positive terms in satellites
	Number of negative terms in satellites
	Number of exclamations and interrogations in the nucleus
	Number of exclamations and interrogations in satellites
	Proportion of positive terms in the nucleus
	Proportion of negative terms in the nucleus
	Proportion of positive terms in satellites
	Proportion of negative terms in satellites
	Proportion of exclamations and interrogations in the nucleus
	Proportion of exclamations and interrogations in satellites

Table 4: Sentence features for polarity classification. The features related to satellites are defined for each specific type of rhetorical relation mentioned in Table 1.

(by normalising by the length of the discourse unit). We also encoded the number and proportion of exclamations and interrogations in the nucleus and satellites. Interrogations and exclamations have been successfully applied in other fine-grained opinion mining scenarios, such as sentiment detection in tweets [47].

Table 4 summarises the considered sentence features. We employed these feature-based representation to build *linear* classifiers (Support Vector Machines or Logistic Regression). Such classifiers base their decision rule on a weighted combination of the feature values, thus bringing the advantage of easily interpretable weights that are assigned to input features in the learning process. This can be seen as an extension of the work presented in Section 3. For instance, the linear combination defined in equation 4 can naturally be learned by these classifiers.

4.2. Experiments

We have evaluated our method on two test collections, containing sentence-level relevance and opinion judgments. The aim of these experiments was to assess the extent to which our RST-proposed based features contribute to a better understanding of text, resulting in improved polarity classification performance.

4.2.1. Data

The English version of the MOAT research collection [44] contains 80 news articles from different sources, and provides 14 topics (describing users' information needs by means of a title and a narrative). All sentences within these documents were annotated by three assessors for relevance and sentiment. We constructed our ground truth for sentence polarity classification by a majority rule: a sentence was regarded as positive (resp. negative) if at least two assessors identified it as positive (resp. negative). This resulted in 596 sentences (179 positive and 417 negative). After removing terms that appear in less than four sentences we obtained 2,218 unique unigrams and 2,812 unique bigrams. We did not apply stemming and we did not remove stop words.

We also performed an evaluation of the FSD collection [45], which contains 294 product reviews from various online sources. The reviews are approximately balanced with respect to domain (covering books, DVDs, electronics, music, and videogames) and overall review sentiment (positive, negative, and neutral). Two annotators assigned sentiment labels to the sentences. The identified sentence-level sentiment is often aligned with the sentiment of the associated reviews, but reviews from all categories contain a substantial fraction of neutral sentences, as well as both positive and negative sentences. The FSD collection includes a total of 2,243 polar sentences: 923 positive sentences, and 1,320 negative sentences. After removing terms that appear in less than four sentences we obtained 1,275 unique unigrams and 1,996 unique bigrams.

4.2.2. Training and Test

We experimented with the linear classifiers of the Liblinear library [48], which supports classification by means of Support Vector Machines (SVMs) and Logistic Regression (LR). We extensively tested these classifiers against the training collection in order to select the best classifier.

Each collection was randomly split into a training set and a test set (75% and 25% of the sentences, respectively) and the classifiers were optimised using 5-fold cross-validation against the training data. For each collection, the classifier that performed the best at training time (in terms of F_1) was subsequently validated with the test set. We repeated this process ten times and we averaged the performance over these ten folds. We measured statistical significance with a paired,

two-sided micro sign test [49], which compares two systems based on their binary decisions and applies the Binomial distribution to compute the p-values under the null hypothesis of equal performance[2].

In both collections, the two-class categorisation problem is slightly unbalanced: the MOAT collection is composed of 179 positive and 417 negative sentences, and the FSD collection contains 923 positive sentences and 1,320 negative ones. Therefore, we tested asymmetric misclassification costs so that positive sentences classified as negative can be penalised more strongly. In order to accomplish this, the parameter C, which penalises all types of errors equally, was tested in the range: $\{1, 2, 3, 5, 10, 50, 100, 1000, 10000, 1000000\}$. The false positive cost, C_{-+}, was always set to C, and the false negative cost, C_{+-}, was set to $C * w$ where w was tested in the same range as C.

4.2.3. Polarity Classification Performance

Tables 5 and 6 show the polarity classification performance on MOAT and FSD, respectively. Vocabulary-based classifiers (unigrams only or both unigrams and bigrams) were regarded as the baselines and we tested the incorporation of various combinations of features into the baseline classifiers. More specifically, we incorporated the Length, Position, RST, and Sentiment RST feature sets detailed in Table 4. Additionally, we ran experiments with all features included (All).

A general trend that can be observed is that our best classifiers tend to have a bias towards negative classifications, which typically show a high recall and a somewhat lower precision. Positive sentences are typically identified with a higher precision than recall. This bias can be attributed to the polarity classes being unequally distributed in the data, which holds especially true for the MOAT collection.

Some of the additional features yield clear performance improvements over solely using vocabulary features, whereas other features do not appear to help much. One trend emerging from our experimental results is the limited extent to which our considered length and positional features contribute to the overall sentence-level polarity classification performance. These features may be useful for detecting opinionated passages in documents [50], but do not have much discriminative power in terms of the polarity of such opinionated passages.

Our considered set of binary RST-based features do not appear to be particularly helpful either. Only small improvements in polarity classification performance can be attributed to these features. These differences were not statistically

[2]Observe that this test considers all label assignments (i.e., 1 means positive sentence and 0 means negative sentence) and, therefore, the result of the test gives a reliable estimation of the overall difference between the systems.

	Positive			Negative			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.439	.220	.293	.729	.882	.798	.686	
+ Length	.393	.283	.297	.725	.846	.781	.666	$\sim$
+ Positional	.399	.272	.323	.730	.828	.776	.663	$\sim$
+ RST	.472	.272	.345	.740	.872	.801	.695	$\sim$
+ Sentiment RST	**.681**	.290	.407	.760	**.943**	**.841**	**.750**	$\gg$
+ All	.530	**.499**	**.514**	**.794**	.814	.804	.721	$>$
Unigrams and bigrams	.615	.218	.322	.741	.943	.830	.728	$\sim$
+ Length	.483	.356	.407	.755	.842	.796	.697	$\ll$
+ Positional	.392	.342	.366	.738	.777	.757	.648	$\ll$
+ RST	.561	.220	.316	.739	.928	.823	.718	$\sim$
+ Sentiment RST	**.744**	.278	.403	.759	**.960**	**.848**	**.758**	$\gg$
+ All	.506	**.562**	**.533**	**.807**	.769	.788	.708	$\sim$

Table 5: Polarity classification results for the MOAT collection, in terms of precision, recall, and F_1 scores for positive and negative sentences. For each vocabulary representation (i.e., unigrams, or unigrams and bigrams), the best performance for each metric is bolded. The symbols $\gg$ and $>$ (resp. $\ll$ and $<$) indicate a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p < 0.01$ and $p < 0.05$, respectively. $\sim$ indicates that the difference was not statistically significant ($p > 0.05$).

significant according to the sign test. Apparently, the presence of particular rhetorical relations per se does not convey much more information than unigrams and bigrams do.

It was only when we used features that combine rhetorical relations with word-level polarity information that performance increased substantially. These Sentiment RST features capture how polar terms are used in a sentence. They allow for differentiation between discourse units, based on their rhetorical roles, when analysing the polarity of these segments. This is in line with our RST-based computation of a sentence's polarity, defined in equation 4 in Section 3.1. These sentence-level polarity classification results validate the observed potential of RST-guided SA in the large-scale polarity ranking task presented in Section 3.

4.2.4. Feature Weights

The feature weights of a linear classifier are indicative of the relevance of each feature. The higher the absolute value of a weight, the more important the feature is for the classifier's decision rule. By analysing the weights, we can gain an understanding of which features are prominent for sentence-level polarity classification.

A proper and direct comparison of the weights can only be done if all features are scaled into the same range. We therefore scaled all features into the range

	Positive			Negative			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.660	.618	.638	.730	.765	.747	.702	
+ Length	.645	.520	.576	.690	**.789**	.736	.674	$\ll$
+ Positional	.672	.622	.646	.735	.776	.755	.710	$>$
+ RST	.669	.607	.637	.728	.778	.752	.705	$\sim$
+ Sentiment RST	**.691**	.674	**.683**	.764	.777	.770	**.734**	$\gg$
+ All	.607	**.738**	.666	**.770**	.646	.703	.685	$<$
Unigrams and bigrams	.680	.587	.630	.723	.796	.758	.707	
+ Length	.662	.459	.542	.674	**.827**	.743	.670	$\ll$
+ Positional	.696	.585	.636	.726	.811	.766	.715	$>$
+ RST	.688	.573	.625	.719	.808	.761	.708	$\sim$
+ Sentiment RST	**.715**	.649	**.681**	**.758**	.809	**.783**	**.741**	$\gg$
+ All	.637	**.705**	.669	.763	.703	.732	.704	$\sim$

Table 6: Polarity classification results for the FSD collection, in terms of precision, recall, and F_1 scores for positive and negative sentences. For each vocabulary representation (i.e., unigrams, or unigrams and bigrams), the best performance for each metric is bolded. The symbols $\gg$ and $>$ (resp. $\ll$ and $<$) indicate a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p < 0.01$ and $p < 0.05$, respectively. $\sim$ indicates that the difference was not statistically significant ($p > 0.05$).

$[0, 1]$. For features thus normalised, a positive feature weight implies that high feature values of the feature are indicative of the sentence being a positive one. Conversely, a negative feature weight suggests that high values of the feature are indicative of the sentence having a negative classification.

The weights in our best performing models exhibit several trends. First, some of the most discriminative vocabulary-based features were unigrams such as *great*, *cool*, or *awesome*, showing that our supervised learning approach can naturally discover and weight sentiment words.

Furthermore, for our best classifiers, which combine vocabulary features with Sentiment RST features, the most discriminative features were the number and proportion of positive and negative terms in the nucleus. In both collections, these four features were among the top ten most discriminative features and the proportion of positive terms was always the most discriminative feature. This highlights the importance of the nucleus of the sentences to understand the polarity of the sentence.

High weights were assigned to some other features combining RST with polarity information, e.g., those associated with attribution, elaboration or enablement relations. Contrast satellites were usually assigned lower weights and, sometimes, even counterbalanced the nucleus weight. For instance, the number of positive

terms got a nucleus weight of 2.070 and a contrast satellite weight of −0.250. This illustrates again the ability of RST relations to facilitate an intricate aggregation of opinion scores from different text segments. A similar outcome was described in Section 3.2.

5. Conclusions and Future Work

We have thoroughly studied the usefulness of RST in various sentiment analysis tasks on different types of information sources. First, we have shown how to infer the sentiment conveyed by a blog post from a small selection of key evaluative sentences. By analysing the rhetorical strucutre of these sentences, we significantly improved on exiting polarity retrieval baselines. The reason of this success lies in aggregating and weighting the sentiment conveyed by text segments with distinct rhetorical roles. For instance, our method accounts for bloggers' apparent tendency of expressing their sentiment in a more apparent fashion in elaborating and attributing text segments, rather than in the core segments of the selected sentences. Additionally, the sentiment conveyed by contrasting text segments is typically inverted in order to better estimate the overall polarity.

In order to validate our findings, we have further studied the potential of RST for sentence-level polarity classification of news and product reviews. Our experimental results show that it is indeed only when we combine rhetorical relations with word-level polarity information that we can obtain clear polarity classification performance improvements. The most valuable features of the polarity classifiers essentially capture the way in which polar terms are used in a sentence. In this respect, the sentence's discourse units and their rhetorical roles are important proxies that can be exploited in SA.

One possible way of further exploiting the potential of RST-guided SA is to represent rhetorical relations in more formal ways, e.g., by using Language Models [18]. Additionally, we would like to explore more efficient and scalable methods for identifying discourse structure of text. Last, we aim to further study the interplay between rhetorical structure of natural language text and its conveyed sentiment, for instance by exploring inter-sentence rhetorical analysis.

Acknowledgements

This work was partially funded by *Secretaría de Estado de Investigación, Desarrollo e Innovación* from the Spanish Government under project TIN2012-33867.

The second author of this paper is partially supported by the Dutch national program COMMIT.

Appendix A. Particle Swarm Optimisation

PSO is a population-based stochastic optimisation technique, inspired by the social behaviour of bird flocking or fish schooling, and included in swarm intelligence techniques. The potential solutions, called *particles*, fly through the problem space following the current optimum particles. The movements of the particles are guided by the best known position of each particle in the search space as well as the entire swarm's best known position. The process is repeated until a satisfactory solution is discovered.

The basic PSO algorithm is summarised in Algorithm 1. Each particle i stores its current position x_i^t , velocity v_i and its best known position pb_i^t at time t. Moreover, the algorithm considers the best known position of the entire swarm (gb^t). We iterated 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.

Algorithm 1 Particle Swarm Optimisation Algorithm

Initialise all particles i with random positions x_i^0 in search space as well as random velocities v_i^0.
Initialise the particle's best known position (pb^0) to its initial position.
Calculate the initial swarm's best known position gb^0 .
repeat
  for all Particle i in the swarm **do**
    Pick random numbers: $rp, rg \in (0, 1)$
    Update the particle's velocity: $v_i^{t+1} = a*v_i^t + b*r_p*(pb_i^t - x_i^t) + c*r_g*(gb^t - x_i^t)$
    Compute the particles new position: $x_i^{t+1} = x_i^t + v_i^{t+1}$
    if $fitness(x_i^{t+1}) < fitness(pb_i^t)$ **then**
      Update the particle's best known position: $pb_i^{t+1} = x_i^{t+1}$
    end if
    if $fitness(pb_i^t) < fitness(gb^t)$ **then**
      Update the swarm's best known position: $gb^{t+1} = pb_i^{t+1}$
    end if
  end for
until Termination criterion is met
return The best known position: gb.

References

[1] E. Metais, Enhancing Information Systems Management with Natural Language Processing Techniques, Data and Knowledge Engineering 41 (2002) 247–272.

[2] A. Hogenboom, F. Hogenboom, F. Frasincar, K. Schouten, O. van der Meer, Semantics-Based Information Extraction for Detecting Economic Events, Multimedia Tools and Applications 64 (2012) 27–52.

[3] X. Wang, X. Jin, M. Chen, K. Zhang, D. Shen, Topic Mining over Asynchronous Text Sequences, IEEE Transactions on Knowledge and Data Engineering 24 (2012) 156–169.

[4] T. Scholz, 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*, Springer, 2013, pp. 1–12.

[5] B. Zhao, Z. Zhang, W. Qian, A. Zhou, Identification of Collective Viewpoints on Microblogs, Data and Knowledge Engineering 87 (2013) 374–393.

[6] H. Mangassarian, H. Artail, A General Framework for Subjective Information Extraction from Unstructured English Text, Data and Knowledge Engineering 62 (2007) 352–367.

[7] R. Schumaker, Y. Zhang, C. Huang, H. Chen, Evaluating Sentiment in Financial News Articles, Decision Support Systems 53 (2012) 458–464.

[8] X. Yu, Y. Liu, X. Huang, A. An, Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain, IEEE Transactions on Knowledge and Data Engineering 24 (2012) 720–734.

[9] B. Pang, L. Lee, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval 2 (2008) 1–135.

[10] B. Liu, Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2012.

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

[12] R. Feldman, Techniques and Applications for Sentiment Analysis, Communications of the ACM 56 (2013) 82–89.

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

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

[15] A. Hogenboom, F. Hogenboom, U. Kaymak, P. Wouters, F. de Jong, Mining Economic Sentiment using Argumentation Structures, in: 7th International Workshop on Web Information Systems Modeling (WISM 2010) at 29th International Conference on Conceptual Modeling (ER 2010), volume 6413 of *Lecture Notes in Computer Science*, Springer, 2010, pp. 200–209.

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

[17] T. Scholz, 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*, Springer, 2013, pp. 272–277.

[18] C. Lioma, B. Larsen, W. Lu, Rhetorical Relations for Information Retrieval, in: 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012), Association for Computing Machinery, 2012, pp. 931–940.

[19] N. Bach, M. Nguyen, 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*, Springer, 2013, pp. 65–76.

[20] A. Ibrahim, 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*, Springer, 2013, pp. 65–76.

[21] T. Nguyen, 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*, Springer, 2013, pp. 278–284.

[22] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, in: Empirical Methods in Natural Language Processing (EMNLP 2002), Association for Computational Linguistics, 2002, pp. 79–86.

[23] A. Devitt, K. Ahmad, 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, 2007, pp. 984–991.

[24] M. Taboada, K. Voll, 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`.

[25] L. Zhou, B. Li, W. Gao, Z. Wei, K. Wong, Unsupervised Discovery of Discourse Relations for Eliminating Intra-sentence Polarity Ambiguities, in: 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011), Association for Computational Linguistics, 2011, pp. 162–171.

[26] C. Zirn, M. Niepert, H. Stuckenschmidt, M. Strube, Fine-Grained Sentiment Analysis with Structural Features, in: 5th International Joint Conference on Natural Language Processing (IJCNLP 2011), Asian Federation of Natural Language Processing, 2011, pp. 336–344.

[27] S. Somasundaran, G. Namata, J. Wiebe, 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), Association for Computational Linguistics, 2009, pp. 170–179.

[28] B. Chardon, F. Benamara, Y. Mathieu, V. Popescu, 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*, Springer, 2013, pp. 25–37.

[29] W. Mann, S. Thompson, Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, Text 8 (1988) 243–281.

[30] J. Chenlo, 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), Association for Computing Machinery, 2011, pp. 365–374.

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

[32] R. Santos, B. He, C. Macdonald, I. Ounis, Integrating Proximity to Subjective Sentences for Blog Opinion Retrieval, in: 31st European Conference on Information Retrieval (ECIR 2009), Springer, 2009, pp. 325–336.

[33] B. He, C. Macdonald, J. He, I. Ounis, An Effective Statistical Approach to Blog Post Opinion Retrieval, in: 17th ACM Conference on Information and Knowledge Management (CIKM 2008), Association for Computing Machinery, 2008, pp. 1063–1072.

[34] B. Pang, 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), Association for Computational Linguistics, 2004, pp. 271–280.

[35] R. Santos, C. Macdonald, R. McCreadie, I. Ounis, I. Soboroff, Information Retrieval on the Blogosphere, Foundations and Trends in Information Retrieval 6 (2012) 1–125.

[36] I. Ounis, C. Macdonald, I. Soboroff, Overview of the TREC 2008 Blog Track, in: 17th Text Retrieval Conference (TREC 2008), National Institute of Standards and Technology, 2008.

[37] L. Polanyi, A. Zaenen, How Okapi came to TREC, in: TREC: Experiments and Evaluation in Information Retrieval, The MIT Press, 2005, pp. 287–299.

[38] T. Wilson, J. Wiebe, 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), Association for Computational Linguistics, 2005, pp. 347–354.

[39] B. He, C. Macdonald, I. Ounis, Ranking Opinionated Blog Posts using OpinionFinder, in: 31st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008), Association for Computing Machinery, 2008, pp. 727–728.

[40] R. Soricut, 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), Association for Computational Linguistics, 2003, pp. 149–156.

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

[42] C. Macdonald, 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`.

[43] J. Parapar, M. Vidal, J. Santos, Finding the Best Parameter Setting: Particle Swarm Optimisation, in: 2nd Spanish Conference on Information Retrieval (CERI 2012), Springer, 2012, pp. 49–60.

[44] Y. Seki, D. Evans, L. Ku, L. Sun, H. Chen, N. Kando, Overview of Multilingual Opinion Analysis Task at NTCIR-7, in: 7th NTCIR Workshop (NTCIR-7), 2008. Available online, `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/`.

[45] O. Tackstrom, R. McDonald, Discovering Fine-Grained Sentiment with Latent Variabel Structured Prediction Models, in: 33rd European Conference on Information Retrieval (ECIR 2011), Springer, 2011, pp. 368–374.

[46] S. Gerani, M. Carman, F. Crestani, Investigating Learning Approaches for Blog Post Opinion Retrieval, in: 31st European Conference on Information Retrieval (ECIR 2009), Springer, 2009, pp. 313–324.

[47] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, Sentiment Analysis of Twitter Data, in: Workshop on Languages in Social Media (LSM 2011), Association for Computational Linguistics, 2011, pp. 30–38.

[48] R. Fan, K. Chang, C. Hsieh, X. Wang, C. Lin, LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research 9 (2008) 1871–1874.

[49] Y. Yang, X. Liu, A Re-examination of Text Categorization Methods, in: 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999), Association for Computing Machinery, 1999, pp. 42–49.

[50] J. Chenlo, D. Losada, A Machine Learning Approach for Subjectivity Classification Based on Positional and Discourse Features, in: Multidisciplinary Information Retrieval, volume 8201 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 17–28.

Biography of the Authors

Jose M. Chenlo is a PhD. student in Computer Science & Artificial Intelligence at the University of Santiago de Compostela (Spain). Jose M. Chenlo got his BSc in Computer Science in 2008. In 2009 he finished a master on "Research on Information Technologies" at the University of Santiago de Compostela (USC). Next, he joined the CiTIUS and he started his research career in the field of Information Retrieval. In 2012 he participated in the Yahoo! Research lab internship program in Barcelona (Spain) working on the Semantic Search Group.

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. student at Erasmus University Rotterdam. He is affiliated with the Econometric Institute, the research center for Business Intelligence at the Erasmus Research Institute of Management, and Erasmus Studio. In his research, Alexander explores the utilization 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.

David E. Losada is an Associate Professor ("Profesor Titular de Universidad") in Computer Science & Artificial Intelligence at the University of Santiago de Compostela (Spain). David E. Losada received his BSc in Computer Science (with honors) in 1997, and his PhD in Computer Science (with honors) in 2001, both from the University of A Coruña (Spain). In 2003, he joined the Univ. of Santiago de Compostela as a senior researcher. He is currently a permanent faculty member and his areas of interest are Information Retrieval, Text Classification, Natural Language Processing and Opinion Mining.