An Empirical Study of Sentence Features for Subjectivity and Polarity Classification

Jose M. Chenlo*, David E. Losada

Centro de Investigación en Tecnoloxías da Información (CITIUS)
Universidad de Santiago de Compostela, Spain

Abstract

While a number of isolated studies have analysed how different sentence features are beneficial in Sentiment Analysis, a complete picture of their effectiveness is still lacking. In this paper we extend and combine the body of empirical evidence regarding sentence subjectivity classification and sentence polarity classification, and provide a comprehensive analysis of the relative importance of each set of features using data from multiple benchmarks. To the best of our knowledge, this is the first study that evaluates a highly diversified set of sentence features for the two main sentiment classification tasks.

Keywords: Sentiment Analysis, Opinion Mining, Sentence-level Analysis, Subjectivity Classification, Polarity Classification

1. Introduction

Sentiment Analysis (SA) –also known as Opinion Mining– is an active and influential research area concerned with automatically extracting subjectivity from natural language text [30, 23, 7, 14]. Sentence-level analysis plays a major role because it permits a fine-grained view of different opinions expressed in text. Moving closer to opinion targets and sentiments on targets facilitates opinion extraction from text that may only contain a few sentences that discuss the topic of interest [23].

A wide range of studies have demonstrated that information provided by some textual features is valuable for sentiment classification. Many types of sentence features have been proposed and tested in the literature: n-grams, part-of-speech

*Corresponding author
Email addresses: josemanuel.gonzalez@usc.es (Jose M. Chenlo), david.losada@usc.es (David E. Losada)

features, location-based features, lexicon-based features, syntactic features, structural or discourse features, just to name a few. However, there is a lack of substantive empirical evidence of their relative effectiveness with different types of texts. Some features have only been tested against product or movie reviews. Other features have only been tried with news datasets. Furthermore, some experiments were performed in a supervised setting while others were performed in an unsupervised setting. The specific evaluation tasks performed are also diverse, e.g., subjectivity classification, polarity classification, opinion summarisation, or polarity ranking. This heterogeneous array of experimental results makes it difficult to understand what features are effective, and when and how they are best used. For instance, some linguistic cues might be beneficial in news articles (due to the nature of the language utilised by journalists) and harmful in product reviews (where customers tend to use a more direct style).

To shed light on these issues, in this paper we aim to provide a comprehensive body of evidence regarding the effectiveness and limitations of different sentence features. Our main contribution is of an empirical nature: the focus of this paper is to report an extensive set of experiments planned to evaluate the relative effectiveness of different sentence features for subjectivity classification and polarity classification. Specifically, we designed a general experimental setting –with two SA tasks and different sources of data– to jointly evaluate features that have otherwise been tested independently. To the best of our knowledge, this is the first study of this kind for SA at sentence level.

Our analysis of results suggests possible use cases of the features considered and gives insights into the potential applicability of every feature set. The experimental outcome also reveals that some features claimed as effective in the literature have little value once the models incorporate other elemental features.

The remainder of this paper is organised as follows. In Section 2 we discuss related work on existing SA approaches and review how different aspects of content are typically involved in such methods. Sections 3 and 4 describe the method and experiments, respectively. In Section 5 we discuss the results of the evaluation and the main lessons learnt about the relative performance of each feature set. Last, in Section 6, we present our conclusions and suggest directions for future research.

2. Related Work

The state-of-the-art in automated SA has been reviewed extensively. Existing surveys [30, 23, 7, 14], which cover all important topics and recent advances, often pay special attention to sentence-level analysis. For instance, Chapter 4 of Liu's book [23] is devoted to sentence subjectivity and sentiment classification. Liu describes different types of sentence features and learning algorithms –supervised and unsupervised– that have been applied for sentiment classification.

Pang and Lee's survey [30] includes a comprehensive discussion of features that have been explored for SA and outlines how opinion extraction problems are often cast as sentence or passage-level classification problems. The reviews done by Cambria et al [7] and Feldman [14] briefly discuss the main research problems related to sentence-level sentiment analysis and some of the techniques used to solve them.

Most of the methods to extract opinions at sentence level are supervised. Pang et al. [29] made one of the first attempts to apply supervised learning. They showed that it is viable to build effective polarity classifiers for movie reviews. In the last decade, many other researchers have applied classification in SA [23, 44]. As usual in Machine Learning, the selection of features is of paramount importance. Some typical features are n-grams, part-of-speech (POS) features, sentiment words features, syntactic patterns, location features, concept-level features, and discourse features. The next paragraphs briefly review how these features have been employed by different researchers.

N-grams. Pang and Lee [31] demonstrated the usefulness of these features for polarity classification of film reviews. Unigram presence features turned out to be the most effective. Other features were considered, including bigrams, POS and location evidence, but none of these provided consistently better performance once unigrams were incorporated. Paltoglou and Thelwall [28] studied and compared different unigram weighting schemes and concluded that some variants of tf/idf are well suited for SA. Their study was done against movie reviews, product reviews and a blog dataset. A new unigram weighting scheme called Delta TFIDF was proposed for SA by Martineau and Finin [27].

POS. Turney showed that adjectives are important indicators of opinions [41], and Yu and Hatzivassiloglou [46] classified subjective sentences based on subsets of adjectives that were manually tagged as positive or negative.

Sentiment Words. Kim and Hovy determined the sentiment orientation of a sentence by multiplying the scores of the sentiment words in the sentence [22]. Qiu et al. presented a self-supervised method that utilises sentiment lexicons to bootstrap training data from a collection of reviews [32]. Taboada et al. [37] exploited a dictionary of sentiment words and phrases with their associated orientations and strength, and incorporated intensification and negation to compute a sentiment score for each document.

Syntactic Patterns. Turney defined in [41] five syntactic patterns to extract opinions from reviews. These patterns turned out to be very effective for sentiment classification in an unsupervised manner and have become reference rules for discovering opinions [23]. Wiebe and Riloff [42] discovered patterns to feed a rule-based method that generates training data for subjectivity classification. The rule-based subjectivity classifier classifies a sentence as subjective if it contains two or more strong subjective clues. Liu and Seneff proposed an approach for extracting adverb-adjective-noun phrases (e.g., *"very nice car"*) based on the clause structure obtained by parsing sentences into a hierarchical representation [24]. Berardi et al. [3] combined POS, Syntactic Patterns and Sentiment Words to extract opinions from sentences that contain an hyperlink. The extracted opinions were employed to estimate whether a given hyperlink is a citation with positive or negative nature. This method was tested against a blog distillation benchmark.

Location. Pang and Lee built polarity classifiers based on sentences from different parts of a movie review (e.g. first sentences, last sentences) [29]. The results obtained showed that the last sentences of a document might be a good indicator of the overall polarity of the review. Beineke et al. proposed several sentiment summarisation approaches [2] and suggested that the first and the last sentences of the reviews are more important for summarising opinions.

Discourse. Taboada et al. [38] demonstrated the importance of general discourse analysis in polarity classification of multi-party meetings. Related to this, Heerschop et al. [19] worked with film reviews and used rhetorical features to determine the importance of every piece of text in the review for polarity classification. By dividing the text into important and less important parts, depending on their rhetorical role according to a sentence-level analysis, they were able to outperform a document level approach based on polarity lexicons. Chenlo et al. also demonstrated the possibilities of discourse analysis at sentence level [9, 11].

Concept-level. The purpose of concept-level SA is to enrich word-level analysis with semantic and affective information associated with natural language opinions [4, 8, 6]. Concept-based approaches can handle multi-word expressions that are related to opinionated concepts (e.g., *"stay out trouble"*). These methods exploit Web ontologies or semantic networks to infer concepts –and the relations among them– from textual documents. Several efforts have been made to construct concept-based affective knowledge bases, such as SenticNet [5]. Many studies of concept-level sentiment analysis are based on extracting opinionated expressions from these knowledge bases. For instance, in [40], a concept-level sentiment dictionary is built by propagating sentiment values based on the assumption that semantically related concepts share common sentiment. Several teams have employed these concept-level resources in SA problems. For instance, Garcia-Moya et al. [15] proposed a novel concept-aware methodology for retrieving product aspects from a collection of free-text customer reviews. However, in classification for SA, there is a lack of empirical evidence on concept-level features and how they compare to others types of features.

In summary, a wide range of features have been independently tested by a large

Paper	Source	Granularity	Size	Task	Features Considered	Best Performing Features
Pang et al. [31]	movie reviews	docs.	1400	2-class (pos/neg)	ng, pos, l	ng
Turney [41]	reviews	docs.	410	2-class (pos/neg)	sp	sp
Kim and Hovy [22]	newswire docs.	sents.	100	3-class (pos/neg/neu)	pos, sw	pos+sw
Pang and Lee [29]	movie reviews	docs.	2000	2-class (pos/neg)	ng, l	ng
Beineke et al. [2]	movie reviews	sents.	2500	summarisation	ng, l	ng+l
Wiebe and Liloff [42]	press articles	sents.	9289	2-class (subj/obj)	pos, sp	pos+sp
Taboada et al. [38]	reviews	docs.	400	2-class (pos/neg)	pos, sw, d	pos+sw+d
Heerschop et al. [19]	movie reviews	docs.	1000	2-class (pos/neg)	sw, d	sw+d
García-Moya et al. [15]	customer reviews	docs	17174	aspect-level summarisation	c	c

Table 1: Main characteristics of some published studies. The table reports the data source, the granularity of the analysis (sentence-level or document-level), the size of the collection, the type of task (e.g., two-class classification or summarisation), the features considered and the best performing features. The feature sets are labelled as follows: N-grams (ng), POS (pos), Sentiment Words (sw), Syntactic Patterns (sp), Location (l), Discourse (d) and Concepts (c).

number of teams –mostly in constrained settings. Table 1 summarises the main characteristics of some of the studies performed in this area. There is not a clear picture of the impact of every feature set and there is little evidence regarding how some features behave with different types of text (e.g., some features were only tried against reviews). There is a need of systematic studies that compare the most meaningful features under uniform conditions. In this paper we fill this gap and we do it for both subjectivity classification and polarity classification.

3. Method

We are concerned with two sentence-level classification problems: 1) subjectivity classification –e.g., subjective vs. objective–, and 2) polarity classification –e.g., positive vs. negative. These categorisation tasks can be performed by automatic classifiers constructed from training data. The characteristics of sentences can be very well encoded as features in a vector representation. These vectors and the corresponding ground truth labels –subjectivity/polarity class assignments– feed 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 corpus. Unigrams and bigrams are valuable to detect specific domain-dependent (opinionated) expressions. The discriminative power of this type of content features has been demonstrated by several opinion mining studies [31, 16]. Our representation is binary because we work at sentence level. Non-binary weighting schemes [27, 28] are more suited to larger chunks of text.

- ***Positional features.*** Positional evidence could be a good guidance for subjectivity or polarity classification. Recent studies indicated that the position of sentences in a document is an important cue in Sentiment Analysis [31, 29, 10]. We included features in our study that encode the absolute position of the sentence within the document (e.g., 2 for the second sentence in the document), and its relative position (i.e., absolute position normalised by the number of sentences in the document).

- ***Part-of-speech features.*** The part-of-speech (POS) of each word can be valuable for analysing sentiments [41]. For each POS tag –e.g., JJ for adjectives– we defined one sentence feature: the count of occurrences of the tag in the sentence. Text was processed by the Stanford Log-linear Part-Of-Speech Tagger[1], which assigns Treebank POS tags [26] (see Table 2).

- ***Syntactic Patterns.*** In addition to sentiment words, many other language compositions express or imply sentiment and opinions [23]. For instance, Turney [41] defined five syntactic patterns to extract opinions from reviews. These patterns have become reference rules for automatically determining opinions [23]. Turney's patterns are sequences of POS tags (see Table 3)

[1] `http://nlp.stanford.edu/downloads/tagger.shtml`

CC	Coordinating conjunction	PRP$	pronoun, possessive
CD	Cardinal Number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	*to*
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNP	Proper noun, singular	VBP	Verb, non-3rd person singular present
NNPS	Proper noun, plural	VBZ	Verb, 3rd person singular present
NNS	Noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoum
POS	Possessive ending	WP$	Possessive wh-pronoum
PRP	Pronoun, personal	WRB	Wh-adverb

Table 2: Penn Treebank Part-Of-Speech (POS) tags.

and we encoded these as binary features (representing the appearance of every pattern in the sentence).

- ***Sentiment Lexicon features.*** These features account for the number of opinionated terms that occur in the sentence. Sentiment words are recognizably a dominant factor in Opinion Mining [30, 23], where many researchers have successfully employed lexicons to reveal opinions, e.g., [22, 20]. We included lexicon-based features as follows. For subjectivity classification, we computed the number of subjective terms in the sentence. For polarity classification, we represented the number of positive and the number of negative terms in the sentence. We also included the proportion of subjective, positive and negative terms, and the number and proportion of exclamations and interrogations. Interrogations and exclamations have been successfully applied in other fine-grained opinion mining scenarios, such as sentiment detection in tweets [1]. The opinionated terms were obtained from the OpinionFinder (OF) [43] sentiment lexicon.

First Word	Second Word	Third Word
JJ	NN or NNS	anything
RB,RBR, or RBS	JJ	not NN nor NNS
JJ	JJ	not NN nor NNS
NN or NNS	JJ	not NN nor NNS
RB,RBR, or RBS	VB,VBD,VBN, or VBG	Anything

Table 3: Patterns of POS tags defined by Turney [41] for extracting opinions.

- ***RST features.*** Rhetorical Structure Theory (RST) [25] is one of the leading discourse theories. This theory explains how texts can be split into segments that are rhetorically related to one another. Within this structure, text segments can be either nuclei or satellites, with nuclei being assumed to be more significant than satellites with respect to understanding and interpreting a text. Many types of relations between text segments exist; the main paper on RST defines 23 types of relations [25]. A satellite may for instance be an elaboration, an explanation or an evaluation on what is explained in a nucleus. Evidence on specific types of satellites can serve as a guidance for the opinion detection process. For example, an attribution relationship could be indicative of subjectivity. We used SPADE (Sentence-level PArsing of DiscoursE) [36], which creates RST trees for individual sentences and we included binary features associated to the appearance of every type of RST relation (see Table 4) in a given sentence. 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 counts of the positive and negative terms that occur in the nucleus and in every type of satellite. In this way, we individually represented the positivity and negativity of the nucleus, and the positivity and negativity of every type of satellite. Again, the representation is sparse because every sentence only contains one (top-level) satellite type. The positive and negative terms were also obtained from the OF [43] sentiment lexicon. We included absolute and relative counts (by normalising by the length of the discourse unit), and the number and proportion of exclamations and interrogations in the nucleus and satellites.

- ***Concept-level features.*** These features account for the opinionated nature of the concepts occurring in the sentence. We measured different dimensional aspects of those concepts, including: Pleasantness, Sensitivity, Aptitude, Attention and the overall opinionated nature. To meet this aim, we used the

SenticNet 2 corpus [5]. This resource contains semantic and affective information for about 14,000 ConceptNet[2] concepts (multiple words). This helps to go beyond a bag of words (BOW) representation. For instance, the concept *"stay out trouble"* has a positive orientation according to *SenticNet*. However, a BOW approach would probably assign a negative polarity score (because of the presence of the word *trouble*). *SenticNet* contains concepts that are expressed with up to 6 words. For subjectivity classification the scores were represented in absolute value because large absolute values are associated to highly opinionated concepts. For polarity classification, we maintained the *SenticNet* scores as they are because the sign of the score is a major guidance for determining polarity.

- ***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.

Table 5 and Table 6 summarise the considered sentence features for subjectivity and polarity classification. We employed these feature-based representations 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.

Although many real-world problems have an inherently non-linear structure, non-linear classifiers hardly provide any advantage in Text Classification [21]. In such high dimensional problems the models are already sufficiently complex and applying non-linear decision boundaries may lead to overfitting [17]. Furthermore, non-linear decision boundaries take substantially longer to train and designing efficient non-linear classifiers of high dimensional data is still challenging [13].

4. Experiments

Our study focuses on assessing the extent to which different types of sentence features contribute to a better subjectivity and polarity classification. To meet this aim, we evaluated our method on four test collections that supply sentence-level annotations:

[2] `http://conceptnet5.media.mit.edu/`

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

- **Multilingual Opinion Analysis Test collection (MOAT)**. The English part of the MOAT research collection [35] contains 80 news articles from different sources, and provides 14 topics (describing users' information needs, with 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 subjectivity and polarity classification by majority: a sentence was regarded as subjective (resp. objective) if at least two assessors labelled it as such; and, similarly, a sentence was regarded as positive (resp. negative) if at least two assessors labelled it as as such. This resulted in 887 subjective sentences –out of a total of 3584 in the test set– and 596 polar sentences (179 positive and 417 negative).

- **Finegrained Sentiment Dataset (FSD)**. The FSD collection [39] contains 294 product reviews from various online sources. The reviews are approximately balanced with respect to domain (books, DVDs, electronics, music, and videogames) and overall review sentiment (positive, negative, and neutral). Two annotators assigned sentiment labels to sentences. The identified sentence-level sentiment is often aligned with the sentiment of the associ-

Set	**Feature**
Vocabulary	Unigrams and bigrams (binary)
Length (4 feat.)	Length of the sentence Length of the nucleus Length of the satellite Length of the document that contains the sentence
Positional (2 feat.)	Absolute position of the sentence in the document Relative position of the sentence in the document
POS (36 feat.)	Number of occurrences of every POS tags (one feature for each POS tag, see Table 2)
Syntactic Patterns (5 feat.)	The presence of a POS syntactic pattern (one binary feature for each pattern defined in Table 3)
Sentiment Lexicon (4 feat.)	Number and proportion of subjective terms in the sentence Number and proportion of exclamations and interrogations in the sentence
RST (15 feat.)	Contains a satellite (binary) Contains specific satellite types (binary)
Concept-Level (10 feat.)	Sum of scores of pleasantness (abs. value) of concepts in the sentence Sum of scores of sensitivity (abs. value) of concepts in the sentence Sum of scores of aptitude (abs. value) of concepts in the sentence Sum of scores of attention (abs. value) of concepts in the sentence Sum of scores of polarity (abs. value) of concepts in the sentence Avg. score of pleasantness (abs. value) of concepts in the sentence Avg. score of sensitivity (abs. value) of concepts in the sentence Avg. score of aptitude (abs. value) of concepts in the sentence Avg. score of attention (abs. value) of concepts in the sentence Avg. score of polarity (abs. value) of concepts in the sentence
Sentiment RST (56 feat.)	Number and proportion of subjective terms in the nucleus Number and proportion of subjective terms in satellites Number and proportion of exclamations and interrogations in the nucleus Number and proportion of exclamations and interrogations in satellites

Table 5: Sentence features for subjectivity classification. The features related to satellites are defined for each specific type of rhetorical relation mentioned in Table 4.

ated 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 2243 polar sentences: 923 positive sentences and 1320 negative sentences.

Set	**Feature**
Vocabulary	Unigrams and bigrams (binary)
Positional (2 feat.)	Length of the sentence
	Length of the nucleus
	Length of the satellite
	Length of the document that contains the sentence
Positional (2 feat.)	Absolute position of the sentence in the document
	Relative position of the sentence in the document
POS (36 feat.)	Number of occurrences of every POS tags (one feature for each POS tag, see Table 2)
Syntactic Patterns (5 feat.)	The presence of a POS syntactic pattern (one binary feature for each pattern defined in Table 3)
Sentiment Lexicon (6 feat.)	Number and proportion of positive terms in the sentence
	Number and proportion of negative terms in the sentence
	Number and proportion of exclamations and interrogations in the sentence
RST (15 feat.)	Contains a satellite (binary)
	Contains specific satellite types (binary)
Concept-Level (10 feat.)	Sum of scores of pleasantness of concepts in the sentence
	Sum of scores of sensitivity of concepts in the sentence
	Sum of scores of aptitude of concepts in the sentence
	Sum of scores of attention of concepts in the sentence
	Sum of scores of polarity of concepts in the sentence
	Avg. score of pleasantness of concepts in the sentence
	Avg. score of sensitivity of concepts in the sentence
	Avg. score of aptitude of concepts in the sentence
	Avg. score of attention of concepts in the sentence
	Avg. score of polarity of concepts in the sentence
Sentiment RST (90 feat.)	Number and proportion of positive terms in the nucleus
	Number and proportion of negative terms in the nucleus
	Number and proportion of positive terms in satellites
	Number and proportion of negative terms in satellites
	Number and proportion of exclamations and interrogations in the nucleus
	Number and proportion of exclamations and interrogations in satellites

Table 6: Sentence features for polarity classification. The features related to satellites are defined for each specific type of rhetorical relation mentioned in Table 4.

- **Multi-Perspective Question Answering dataset (MPQA)**. This corpus contains news articles manually annotated using an annotation scheme for opinions and other private states (i.e., beliefs, emotions, sentiments or speculations). We followed existing practice [34, 33] that applies annotation patterns to label sentences as either subjective or objective[3]. The same patterns can be easily extended for assigning positive and negative labels. After applying these patterns to the sentence collection we obtained 7333 subjective sentences –out of a total of 15802– and 4881 polar sentences (1626 positive and 3255 negative).

- **Pang & Lee subjectivity dataset (PL)**. PL is an automatically labelled sentence corpus [29]. To gather subjective sentences (or phrases), 5000 review snippets were crawled from a popular film reviews site[4] (e.g., "bold, imaginative, and impossible to resist"). Sentences estimated as objective were obtained from plot summaries of the Internet Movie Database[5].

MOAT and MPQA are suitable for both subjectivity and polarity classification, while PL and FSD can only be used for subjectivity and polarity classification[6], respectively. The main statistics of these collections are reported in Table 7. Observe that the dimension of the feature vector is equal to the number of unigrams –or unigrams and bigrams– (see Table 7) plus the rest of features presented in Table 5 and Table 6[7]. For instance, the Unigram+All subjectivity classifier constructed from MOAT has 2350 features (2218 unigram features + 132 additional features).

4.1. Training and Test

We experimented with the linear classifiers of the Liblinear library [12]: Support Vector Machines (SVMs) and Logistic Regression (LR). Each collection was randomly split into a training set and a test set (75% and 25% of the sentences, respectively). The classifiers were optimised by applying 5-fold cross-validation against the training data. For each collection, the classifier that performed the best (in terms of F_1) was subsequently validated against the test data. This 75% -25% random splitting process was repeated 10 times and we report the average performance obtained over these then runs[8]. We measured statistical significance with

[3] For instance, a sentence that contains a phrase labelled as highly subjective is regarded as a subjective sentence.

[4] `www.rottentomatoes.com`

[5] `www.imdb.com`

[6] PL only contains subjectivity labels and FSD only contains polarity labels for some subjective sentences.

[7] We ran experiments with the features standardised and found no difference with respect to the experiments with no-standardised features. The results reported refer to the non-standardised case.

[8] This reduces variance of the performance results and makes the comparison less dependent on the specific test split.

Subjectivity				
dataset	# subjective sentences	# objective sentences	#unique unigrams	#unique bigrams
MOAT	887	2697	2218	2812
MPQA	7333	8469	6463	9203
PL	5000	5000	4948	9103
Polarity				
dataset	# subjective sentences	# objective sentences	#unique unigrams	#unique bigrams
MOAT	179	417	2218	2812
MPQA	1626	3255	6463	9203
FSD	923	1320	1275	1996

Table 7: Test collections for experimentation in subjectivity and polarity classification. The tables include the number of unique unigrams and bigrams after pre-processing. We did not apply stemming and we did not remove common words. We only removed terms that appeared in less than four sentences.

a paired, two-sided micro sign test [45]. Rather than using the paired F_1 values, this test compares two systems based on all their binary decisions (sentence-class assignments in our case) and applies the Binomial distribution to compute the p-values under the null hypothesis of equal performance.

In most collections, the two-class categorisation problem is unbalanced: fewer subjective sentences than objective sentences, or fewer positive sentences than negative ones. Therefore, we tested asymmetric misclassification costs so that subjective sentences classified as objective (or positive sentences classified as negative) can be penalised more strongly[9].

4.2. Subjectivity Classification Performance

In Table 8, Table 9 and Table 10 we report the subjectivity classification performance achieved on MOAT, MPQA and PL, respectively. Vocabulary-based classifiers (unigrams only, or unigrams combined with bigrams) were regarded as baselines and we incorporated various combinations of features into the baseline classifiers: Length, Position, POS tags, POS syntactic Patterns, Sentiment Lexicon, Concept-level, RST, and Sentiment RST (see Table 5). Additionally, we ran experiments with all features included (All).

[9] The SVM or LR 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.

	Subjective			Objective			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.5207	.4295	.4707	.8185	.8667	.8419	.7565	
+ Length	.4933	.4907	.4920	.8287	.8301	.8294	.7446	≪
+ Position	.5046	.5111	.5078	.8345	.8309	.8327	.7503	∼
+ POS	.5282	.4362	.4778	.8205	.8687	.8439	.7597	∼
+ Synt. Patterns	.5387	.4286	.4774	.8198	**.8763**	**.8471**	.7635	≫
+ Sentiment Lexicon	**.5401**	.4566	.4949	.8259	.8690	.8469	**.7650**	≫
+ Concept-level	.5302	.4588	.4919	.8255	.8630	.8438	.7611	>
+ RST	.5305	.4242	.4714	.8182	.8735	.8449	.7602	∼
+ Sentiment RST	.5316	.4814	.5053	.8306	.8570	.8436	.7623	>
+ All	.5171	**.5426**	**.5295**	**.8432**	.8293	.8362	.7570	∼
Uni and bigrams	.5802	.3577	.4426	.8083	.9128	.8574	.7728	
+ Length	.5041	.4371	.4682	.8184	.8551	.8363	.7497	≪
+ Position	.5332	.4632	.4957	.8268	.8633	.8447	.7625	≪
+ POS	.5864	.3639	.4491	.8099	**.9135**	.8586	.7750	∼
+ Synt. Patterns	.5747	.3546	.4386	.8074	.9116	.8563	.7712	∼
+ Sentiment Lexicon	**.589**5	.3941	.4724	.8163	.9075	**.8595**	**.7781**	>
+ Concept-level	.5645	.3994	.4678	.8157	.8962	.8541	.7709	∼
+ RST	.5587	.3794	.4519	.8113	.8990	.8529	.7680	<
+ Sentiment RST	.5698	.4326	.4918	.8231	.8899	.8552	.7746	∼
+ All	.5042	**.5328**	**.5181**	**.8395**	.8234	.8314	.7502	≪

Table 8: Subjectivity classification results for the MOAT collection, in terms of precision, recall, and F_1 scores for subjective and objective sentences. For each vocabulary representation (i.e., unigrams, or unigrams and bigrams), the best performance for each metric is bolded. The symbol $\gg$ (resp. $\ll$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p \leq .01$. The symbol $>$ (resp. $<$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baseline with $0.1 < p \leq .05$. $\sim$ indicates that the difference was not statistically significant ($p > .05$).

The results reveal the following trends. Length features do not contribute to discriminate between objective and subjective sentences. Syntactic Patterns seem to be slightly beneficial when used on top of unigram representations. But these linguistic cues do not help in combination with bigrams. This indicates that bigrams are already capturing some structural aspects of subjective sentences.

POS features are valuable: in MPQA and PL they led to statistical significant improvements over the baselines and, in MOAT, performance remained roughly the same. This confirms the usefulness of counting POS labels to detect subjective content.

Positional features seem to work particularly well for discovering subjective

	Subjective			Objective			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.7172	.7222	.7197	.7597	.7552	.7574	.7399	
+ Length	.6683	.7007	.6841	.7315	.7010	.7159	.7009	$\ll$
+ Position	.7311	.7564	.7435	.7841	**.7608**	**.7723**	**.7588**	$\gg$
+ POS	.7217	.7248	.7232	.7625	.7597	.7611	.7436	$\gg$
+ Synt. Patterns	.7170	.7268	.7219	.7623	.7533	.7578	.7410	$\sim$
+ Sentiment Lexicon	.7250	.7383	.7316	.7714	.7592	.7653	.7495	$\gg$
+ Concept-level	**.7350**	.7202	.7275	.7635	.7767	.7700	.7506	$\gg$
+ RST	.7150	.7272	.721	.7620	.7507	.7563	.7399	$\sim$
+ Sentiment RST	.7236	.7370	.7302	.7702	.7579	.7640	.7483	$\gg$
+ All	.7262	**.7673**	**.7462**	**.7897**	.7512	.7700	.7587	$\gg$
Uni and bigrams	.7226	.7069	.7147	.7526	.7666	.7595	.7390	
+ Length	.6756	.7126	.6936	.7407	.7058	.7228	.7089	$\ll$
+ Position	.7248	**.7551**	.7396	.7816	.7534	.7672	.7542	$\gg$
+ POS	.7234	.7135	.7184	.7565	.7655	.7610	.7414	$>$
+ Synt. Patterns	.7191	.7126	.7158	.7548	.7607	.7577	.7384	$\sim$
+ Sentiment Lexicon	.7323	.7212	.7267	.7634	.7733	.7683	.7492	$\gg$
+ Concept-level	.7302	.7139	.7220	.7586	.7731	.7658	.7458	$\gg$
+ RST	.7188	.7123	.7155	.7545	.7604	.7574	.7382	$\sim$
+ Sentiment RST	.7250	.7254	.7252	.7638	.7634	.7636	.7458	$\gg$
+ All	**.7397**	.7439	**.7418**	**.7787**	**.7749**	**.7768**	**.7606**	$\gg$

Table 9: Subjectivity classification results for the MPQA collection, in terms of precision, recall, and F_1 scores for subjective and objective sentences. For each vocabulary representation (i.e., unigrams, or unigrams and bigrams), the best performance for each metric is bolded. The symbol $\gg$ (resp. $\ll$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p \leq .01$. The symbol $>$ (resp. $<$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baseline with $0.1 < p \leq .05$. $\sim$ indicates that the difference was not statistically significant ($p > .05$).

content. Where available[10], positional information helped to improve recall of subjective sentences. However, its ability to classify objective sentences seems to be limited. This might indicate a tendency of using subjective sentences in specific parts of the document, e.g., in the end of the document as a conclusion.

Binary RST-based features did not work well. Apparently, the presence of particular rhetorical relations per se does not convey much more information than unigrams and bigrams do.

The best performing combination was the one that included the sentiment lexicon features. It was the only feature set able to statistically improve the baselines in

[10] Observe that we do not have positional information in the PL collection.

	Subjective			Objective			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.8939	.8910	.8924	.8916	.8944	.8930	.8927	
+ Length	.8614	.8940	.8774	.8901	.8565	.8730	.8752	$\ll$
+ Position	–	–	–	–	–	–	–	
+ POS	**.9007**	**.9008**	**.9007**	**.9010**	**.9009**	**.9009**	**.9008**	$\gg$
+ Synt. Patterns	.8969	.8955	.8962	.8959	.8973	.8966	.8964	$\gg$
+ Sentiment Lexicon	.8926	.8995	.8960	.8989	.8920	.8954	.8958	$\gg$
+ Concept-level	.8919	.8938	.8928	.8938	.8919	.8928	.8928	$\sim$
+ RST	.8934	.8910	.8922	.8915	.8939	.8927	.8924	$\sim$
+ Sentiment RST	.8903	.9004	.8953	.8995	.8892	.8943	.8948	$\sim$
+ All	.8876	.9005	.8940	.8993	.8862	.8927	.8934	$\sim$
Uni and bigrams	.9043	.8942	.8992	.8956	.9055	.9005	.8999	
+ Length	.8829	.8811	.8820	.8816	.8834	.8825	.8822	$\ll$
+ Position	–	–	–	–	–	–	–	
+ POS	**.9099**	.8945	**.9021**	.8965	**.9116**	**.9040**	**.9031**	$>$
+ Synt. Patterns	.9047	.8930	.8988	.8946	.9062	.9004	.8996	$\sim$
+ Sentiment Lexicon	.9016	.8964	.899	.8973	.9024	.8998	.8994	$\sim$
+ Concept-level	.9069	.8916	.8992	.8937	.9087	.9011	.9002	$\sim$
+ RST	.9054	.8888	.8970	.8910	.9073	.8991	.8980	$\ll$
+ Sentiment RST	.9034	.8916	.8975	.8932	.9049	.8990	.8982	$\sim$
+ All	.8999	**.9026**	.9012	.9025	.8999	.9012	.9012	$\sim$

Table 10: Subjectivity classification results for the PL collection, in terms of precision, recall, and F_1 scores for subjective and objective sentences. For each vocabulary representation (i.e., unigrams, or unigrams and bigrams), the best performance for each metric is bolded. The symbol $\gg$ (resp. $\ll$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p \leq .01$. The symbol $>$ (resp. $<$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baseline with $0.1 < p \leq .05$. $\sim$ indicates that the difference was not statistically significant ($p > .05$).

all situations across the different test sets. Combining unigrams or bigrams with a sentiment lexicon is a way to account for both general purpose opinion expressions and domain-specific opinion expressions. This led to robust subjectivity classifiers.

Concept-level features led to some statistically significant improvements but they look inferior to Sentiment lexicon features. This suggests that multiword expressions that represent concepts are not needed for discovering subjectivity (opinionated single-word terms are enough).

Sentiment RST features, which weight opinionated terms within the RST spans of the sentences, led to modest improvements over the baselines. These improvements were inferior to those found with Sentiment Lexicon features. This suggests that Sentiment RST features are not more discriminative than pure lexicon-based features for subjectivity classification.

Finally, when combining all features into a single classifier we obtained a good classifier in terms of recall of subjective sentences but recall of objective sentences tended to fall. This led to classification performance that was sometimes worse than the baseline's performance (e.g., in MOAT, all features combined led to performance decreases that were statistically significant).

4.3. Polarity Classification Performance

In Table 11, Table 12, and Table 13 we report the polarity classification performance on MOAT, MPQA and FSD, respectively. Again, vocabulary-based classifiers were regarded as baselines and we tested the incorporation of various combinations of features into the baseline classifiers.

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.

One trend emerging from the experiments is the limited extent to which our considered length, positional, POS, POS linguistic patterns and RST features contribute to the overall sentence-level polarity classification performance. Some of these features were useful for detecting opinionated passages (see Section 4.2), but do not have much discriminative power in terms of the polarity of such opinionated passages. For instance, position and POS features were indicative of subjectivity but do not help here to estimate polarity. This makes sense because the position of a sentence could arguably be indicative of subjectivity (e.g., a news article might begin with factual content) but it is hardly a polarity cue. Similarly, some POS features, e.g., the number of adjectives, are often indicative of subjectivity but do not reveal by themselves the orientation of the sentiments.

One of the best performing combinations was again the one that includes the sentiment lexicon features. In all cases, this configuration led to significant improvements with respect to the baselines. However, Sentiment RST was the only feature set whose inclusion into the baseline led to improvements with p-value always lower than .01. Concept-level features also seem to give an added value compared to the baseline classifiers. Although still inferior to Sentiment Lexicon features, concept-level features led to improvements over the baselines that were statistically significant in most of the cases. This suggests that they could play a role in larger collections, with massive occurrences of multiword expressions, or for more elaborated tasks. For instance, concepts were successfully employed for detecting opinions about specific product aspects in customer reviews [15] .

	Positive			Negative			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.4389	.2200	.2931	.7289	.8818	.7981	.6859	
+ Length	.3933	.2381	.2966	.7253	.8456	.7808	.6658	∼
+ Position	.3987	.2721	.3235	.7300	.8275	.7757	.6631	∼
+ POS	.4703	.2517	.3279	.7368	.8808	.8024	.6946	∼
+ Synt. Patterns	.5053	.2154	.3020	.7343	.9113	.8133	.7054	∼
+ Sentiment Lexicon	.6505	.3039	.4143	.7609	.9314	.8376	.7456	≫
+ Concept-level	.5385	.2381	.3302	.7405	.9142	.8182	.7141	≫
+ RST	.4724	.2721	.3453	.7403	.8723	.8009	.6946	∼
+ Sentiment RST	**.6809**	.2902	.4070	.7596	**.9428**	**.8413**	.7497	≫
+ All	.5910	**.5079**	**.5463**	**.8047**	.8522	.8278	**.7503**	≫
Uni and bigrams	.6154	.2177	.3216	.7414	.9428	.8301	.7282	
+ Length	.4829	.3515	.4069	.7553	.8418	.7962	.6966	≪
+ Position	.3922	.3424	.3656	.7376	.7769	.7567	.6483	≪
+ POS	.4976	.2381	.3221	.7373	.8990	.8102	.7034	≪
+ Synt. Patterns	.4625	.2517	.3260	.7360	.8770	.8003	.6919	≪
+ Sentiment Lexicon	.6634	.3084	.4211	.7626	.9342	.8397	.7490	>
+ Concept-level	.5862	.1927	.2901	.7353	.9428	.8262	.7208	∼
+ RST	.5607	.2200	.3160	.7388	.9276	.8225	.7181	∼
+ Sentiment RST	**.7439**	.2766	.4033	.7594	**.9600**	**.8480**	**.7577**	≫
+ All	.5761	**.5578**	**.5668**	**.8166**	.8275	.8220	.7477	∼

Table 11: 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 symbol ≫ (resp. ≪) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p \leq .01$. The symbol > (resp. <) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baseline with $0.1 < p \leq .05$. ∼ indicates that the difference was not statistically significant ($p > .05$).

Sentiment RST features help to differentiate between discourse units, based on their rhetorical roles, when analysing the polarity of these segments. This yielded to polarity classifiers that were slightly more robust than those constructed from structure-unaware features (i.e., Sentiment Lexicon). These sentence-level polarity classification results validate the observed potential of RST-guided Sentiment Analysis in the large-scale polarity ranking of blog posts [9].

Finally, the combination of all features worked well, but was inferior to both Sentiment Lexicon and Sentiment RST.

	Positive			Negative			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.6570	.5824	.6175	.8021	.8477	.8243	.7592	
+ Length	.6585	.5893	.6220	.8046	.8470	.8253	.7610	$\sim$
+ Position	.6405	.5996	.6194	.8057	.8315	.8184	.7541	$\sim$
+ POS	.6563	.5854	.6188	.8030	.8465	.8242	.7593	$\sim$
+ Synt. Patterns	.6636	.5824	.6204	.8029	.8521	.8268	.7621	$\sim$
+ Sentiment Lexicon	**.6973**	.6554	.6757	.8325	.8575	.8448	.7901	$\gg$
+ Concept-level	.6591	.6212	.6396	.8156	.8391	.8272	.7664	$>$
+ RST	.6554	.5822	.6166	.8018	.8467	.8236	.7584	$\sim$
+ Sentiment RST	.6881	.6541	.6707	.8310	.8515	.8411	.7857	$\gg$
+ All	.6934	**.6645**	**.6786**	**.8354**	**.8529**	**.8441**	**.7900**	$\gg$
Uni and bigrams	.6770	.5463	.6047	.7928	.8695	.8294	.7616	
+ Length	.6771	.5542	.6095	.7953	.8676	.8299	.7630	$\sim$
+ Position	.6598	.5704	.6119	.7985	.8527	.8247	.7585	$\sim$
+ POS	.6665	.5389	.5959	.7893	.8649	.8254	.7561	$\ll$
+ Synt. Patterns	.6743	.5446	.6025	.7920	.8682	.8284	.7602	$\sim$
+ Sentiment Lexicon	**.7171**	.6362	**.6742**	.8276	**.8743**	**.8503**	**.7948**	$\gg$
+ Concept-level	.6835	.5750	.6246	.8028	.8667	.8335	.7693	$>$
+ RST	.6707	.5473	.6027	.7924	.8654	.8273	.7593	$\sim$
+ Sentiment RST	.7114	.6310	.6688	.8251	.8718	.8478	.7915	$\gg$
+All	.6953	**.6502**	.6720	**.8303**	.8573	.8436	.7882	$\gg$

Table 12: Polarity classification results for the MPQA 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 symbol $\gg$ (resp. $\ll$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p \leq .01$. The symbol $>$ (resp. $<$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baseline with $0.1 < p \leq .05$. $\sim$ indicates that the difference was not statistically significant ($p > .05$).

5. Discussion

In the present work, we found that some sentence features can potentially lead to better and more reliable Sentiment Analysis classifiers:

- Among all features tested, length and binary RST-based features did not give any added value for subjectivity or polarity classification. These features hardly improved over the baselines and, often, led to performance decreases. We conclude that the presence of particular rhetorical relations or the length of the sentence are not indicative of subjectivity or polarity.

- Positional features worked well as a recall-oriented mechanism for detecting subjective sentences. We recommend injecting positional information for

	Positive			Negative			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.6596	.6175	.6379	.7302	.7647	.7471	.7021	
+ Length	.6451	.5195	.5755	.6897	**.7889**	.7360	.6745	≪
+ Position	.6720	.6217	.6459	.7352	.7758	.7550	.7104	>
+ POS	.6740	.6221	.6470	.7359	.7777	.7562	.7116	>
+ Synt. Patterns	.6747	.6389	.6563	.7434	.7724	.7576	.7157	≫
+ Sentiment Lexicon	**.6936**	.6717	**.6825**	.7630	.7808	**.7718**	**.7345**	≫
+ Concept-level	.6809	.6242	.6513	.7385	.7839	.7605	.7161	≫
+ RST	.6690	.6074	.6367	.7285	.7780	.7524	.7055	∼
+ Sentiment RST	.6911	.6742	**.6825**	**.7636**	.7774	.7704	.7336	≫
+ All	.6245	**.7348**	.6752	**.7747**	.6737	.7207	.6996	∼
Uni and bigrams	.6801	.5872	.6302	.7231	.7960	.7578	.7073	
+ Length	.6618	.4590	.5421	.6742	.8268	.7427	.6705	≪
+ Position	.6958	.5855	.6359	.7260	.8109	.7661	.7152	>
+ POS	.6883	.5792	.6291	.7218	.8063	.7617	.7098	∼
+ Synt. Patterns	.6949	.5792	.6318	.7233	.8122	.7652	.7132	∼
+ Sentiment Lexicon	.7149	.6578	**.6852**	.7614	.8063	**.7832**	**.7432**	≫
+ Concept-level	.6971	.5939	.6414	.7296	**.8094**	.7674	.7179	≫
+ RST	.6878	.5734	.6254	.7194	.8078	.7610	.7082	∼
+ Sentiment RST	**.7153**	.6494	.6808	.7576	.8091	.7825	.7412	≫
+ All	.6297	**.7385**	.6798	**.7786**	.6793	.7256	.7045	∼

Table 13: 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 symbol ≫ (resp. ≪) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p \leq .01$. The symbol > (resp. <) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baseline with $0.1 < p \leq .05$. ∼ indicates that the difference was not statistically significant ($p > .05$).

tasks where discovering most of the subjective content (at the expense of precision) is crucial.

- POS features were valuable for subjectivity classification. This confirms the usefulness of counting POS labels to detect subjective content [18, 41]. Nevertheless, a score based on counting sentiment terms from a general-purpose vocabulary, i.e., Sentiment Lexicon, was at least as effective as accounting for POS labels. In the light of these results, we conclude that POS tagging, which takes extra processing time, is not worthwhile for subjectivity classification. Furthermore, POS features, e.g., the number of adjectives or adverbs, did not help to distinguish positive and negative sentences.

- Syntactic patterns were only beneficial when the baseline classifier handled unigrams representations. With bigrams, syntactic patterns did not give any added value. In the literature [41, 23], these syntactic patterns were useful for sentiment classification using unsupervised learning, e.g., to extract potential sentiment words from phrases and perform sentiment classification. Our results demonstrate that a simple bigram representation together with some training data is reasonably effective and does not require POS pattern matching.

- Sentiment Lexicon features were robust for both subjectivity and polarity classification. Lexicon-based methods have been shown to be deficient for coarse-grained Sentiment Analysis tasks, e.g., document polarity ranking [19, 9], where the flow of sentiments and the existence of conflicting opinions demand more evolved technology. However, our results demonstrate that lexicon-based features consistently give high performance in sentence subjectivity classification and sentence polarity classification. These fine-grained tasks seem to fit well with primitive counting methods based on lexicons. Observe also that we combined unigram/bigram features with lexicon-based features. In this way, the classifier takes into account not only opinion scores computed from the lexicon but also domain-specific opinion expressions (unigram/bigram features are weighted in a supervised manner). Combining these two factors yielded very effective subjectivity and polarity classifiers.

- Concept-level features are somehow promising for subjectivity and polarity classification. Although including them leads to improvements over standard BOW classifiers, the resulting performance is inferior to the performance obtained with other feature sets (e.g., Sentiment Lexicons). Still, concept-level features could be included into the models in more elaborated ways. For instance, exploiting relationships between concepts, or combining concepts and RST information. This will be subject to further research.

- For subjectivity estimation, Sentiment RST features were slightly inferior to pure lexicon-based features. However, Sentiment RST features were very robust for polarity classification. Sentiment RST features can capture the specific relations between the different parts of the sentences and weight the polar terms accordingly. For instance, the contrast statement presented in the sentence *"The film was awful but it was nice going with her"* cannot be merely solved with lexicon-based approaches.

Overall, classifying sentences based on sentiment lexicon scores and unigrams-bigrams is an effective and safe choice for subjectivity classification. In polarity classification, sentiment lexicon together with unigrams/bigrams are also quite accurate. However, Sentiment RST features are slightly more robust than Sentiment Lexicon features and, therefore, unigrams/bigrams+Sentiment RST seems like a sensible choice for polarity classification at sentence level. Additionally, other sentence features might be considered when some class or performance measure needs to be accentuated (e.g., positional features for an application domain that demands high recall of subjective sentences). And concept-level features permit a semantic analysis. This could be interesting for certain Sentiment Analysis tasks.

Many research studies do Sentiment Analysis at document level (e.g., with full reviews or blog posts), where it is relatively easy to build a large-scale benchmark. At sentence or passage level, it is more difficult to have access to collections with large numbers of labelled sentences. Although we experimented with as many testbeds as possible, our analysis can still be broadened to other domains and sources of data. Our study will be therefore subject to expansion as more data become available.

Our study is constrained to a supervised setting. But a set of annotated sentences is not always available. We do not claim that the relative merits of the feature sets would hold with no training data. Related to this, it would be interesting to study the relative behaviour of the proposed features with little –or in the absence of– training data. Another intriguing line of future work would be to study how these features support transfer learning (i.e., training and test data come from different domains).

6. Conclusions

In this paper we have presented a systematic study of different sentence features for SA. We explored the behaviour of these features against different benchmarks and tasks (subjectivity classification and polarity classification of sentences). Our results reveal interesting tendencies and, overall, the study gives substantive empirical evidence to those interested in sentence-level SA.

We found that unigrams/bigrams combined with sentiment lexicon features consistently give good performance for subjectivity classification. In the literature, other features –e.g., POS labels– were shown to be effective for subjectivity classification but our experiments suggest that once unigrams/bigrams and sentiment lexicon features are incorporated the effect of any other feature is negligible.

Regarding polarity classification, unigrams/bigrams combined with sentiment lexicon features were also effective but the most robust classifier was obtained with unigrams/bigrams combined with sentiment RST features. This shows that these linguistic-based features are valuable to properly understand the sentiment conveyed in sentences.

Acknowledgments

This work was supported by the *"Ministerio de Economía y Competitividad"* of the Goverment of Spain under the research project TIN2012-33867. We thank David Elsweiler and Alvaro Barreiro for their useful comments and suggestions.

References

[1] 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.

[2] P. Beineke, T. Hastie, C. Manning, S. Vaithyanathan, Exploring sentiment summarization, in: Proc. AAAI Spring Symposium on Exploring Attitude and Affect in Text Theories and Applications, pp. 12–15.

[3] G. Berardi, A. Esuli, F. Sebastiani, F. Silvestri, Endorsements and rebuttals in blog distillation, Information Sciences 249 (2013) 38 – 47.

[4] E. Cambria, An introduction to concept-level sentiment analysis., in: F. Castro, A.F. Gelbukh, M. Gonzlez (Eds.), 12th Mexican International Conference on Artificial Intelligence, volume 8266 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 478–483.

[5] E. Cambria, C. Havasi, A. Hussain, Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis., in: G.M. Youngblood, P.M. McCarthy (Eds.), The Florida Artificial Intelligence Research Society Conference, AAAI Press, 2012.

[6] E. Cambria, B. Schuller, B. Liu, H. Wang, C. Havasi, Knowledge-based approaches to concept-level sentiment analysis, IEEE Intelligent Systems 28 (2013) 12–14.

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

[8] E. Cambria, B. Schuller, Y. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis, IEEE Intelligent Systems 28 (2013) 15–21.

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

[10] 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.

[11] 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.

[12] 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.

[13] M. Fauvel, J. Chanussot, J. Benediktsson, A. Villa, Parsimonious mahalanobis kernel for the classification of high dimensional data, Pattern Recognition 46 (2013) 845 – 854.

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

[15] L. García-Moya, H. Anaya-Sánchez, R.B. Llavori, Retrieving product features and opinions from customer reviews., IEEE Intelligent Systems 28 (2013) 19–27.

[16] 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.

[17] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.

[18] V. Hatzivassiloglou, J.M. Wiebe, Effects of adjective orientation and gradability on sentence subjectivity, in: Proceedings of the 18th Conference on Computational Linguistics - Volume 1, COLING '00, Association for Computational Linguistics, Stroudsburg, PA, USA, 2000, pp. 299–305.

[19] 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.

[20] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, ACM, New York, NY, USA, 2004, pp. 168–177.

[21] T. Joachims, Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms, Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[22] S.M. Kim, E. Hovy, Determining the sentiment of opinions, in: Proceedings of the 20th International Conference on Computational Linguistics, COLING '04, Association for Computational Linguistics, Stroudsburg, PA, USA, 2004.

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

[24] J. Liu, S. Seneff, Review sentiment scoring via a parse-and-paraphrase paradigm, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1, EMNLP '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 161–169.

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

[26] M.P. Marcus, B. Santorini, M.A. Marcinkiewicz, Building a Large Annotated Corpus of English: The Penn Treebank, Computational Linguistics 19 (1993) 313–330.

[27] J. Martineau, T. Finin, Delta tfidf: An improved feature space for sentiment analysis., in: E. Adar, M. Hurst, T. Finin, N.S. Glance, N. Nicolov, B.L. Tseng (Eds.), Proceedings of the 3rd AAAI Conference on Weblogs and Social Media (ICWSM), The AAAI Press, 2009.

[28] G. Paltoglou, M. Thelwall, A study of information retrieval weighting schemes for sentiment analysis, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 1386–1395.

[29] 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.

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

[31] 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.

[32] L. Qiu, W. Zhang, C. Hu, K. Zhao, Selc: A self-supervised model for sentiment classification, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, ACM, New York, NY, USA, 2009, pp. 929–936.

[33] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP '03, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 105–112.

[34] E. Riloff, J. Wiebe, T. Wilson, Learning subjective nouns using extraction pattern bootstrapping, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, CONLL '03, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 25–32.

[35] 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). Available online, `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/`.

[36] 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.

[37] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Computational Linguistics 37 (2011) 267–307.

[38] 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`.

[39] 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.

[40] A.C.R. Tsai, C.E. Wu, R.T.H. Tsai, J.Y.J. Hsu, Building a concept-level sentiment dictionary based on commonsense knowledge, IEEE Intelligent Systems 28 (2013) 22–30.

[41] P.D. Turney, Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 417–424.

[42] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, in: Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing'05, Springer-Verlag, Berlin, Heidelberg, 2005, pp. 486–497.

[43] 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.

[44] R. Xia, C. Zong, S. Li, Ensemble of feature sets and classification algorithms for sentiment classification, Information Sciences 181 (2011) 1138–1152.

[45] 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.

[46] H. Yu, V. Hatzivassiloglou, Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, in: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP '03, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 129–136.

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.

- 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.