Interpreting SentiWordNet for Opinion Classification

Horacio Saggion[α], **Adam Funk**[β]

[α]Department of Information and Communication Technologies (DTIC)
Universitat Pompeu Fabra
Campus de la Comunicacio Poble Nou
C/Tanger, 122-140
Barcelona, Spain
[β]Department of Computer Science
University of Sheffield
211 Portobello Street - Sheffield, England, UK
H.Saggion@dcs.shef.ac.uk;A.Funk@dcs.shef.ac.uk

Abstract

We describe a set of tools, resources, and experiments for opinion classification in business-related datasources in two languages. In particular we concentrate on SentiWordNet text interpretation to produce word, sentence, and text-based sentiment features for opinion classification. We achieve good results in experiments using supervised learning machine over syntactic and sentiment-based features. We also show preliminary experiments where the use of summaries before opinion classification provides competitive advantage over the use of full documents.

1. Introduction

Public opinion has a great impact on company and government decision making (Saggion and Funk, 2009). For example, many UK political figures had to apology after public complaints and indignation in the recent *Parliamentary expenses scandal*; while in another case, a supermarket had to stop buying lingerie products from a notorious model, after customers' complaints about her inappropriate behavior were posted in blogs and Internet fora. The Web has become an important source of information, in the field of business intelligence, business analysts are turning their eyes on the web in order to obtain factual as well as more subtle and subjective information (opinions) on companies and products. However without appropriate tools, identifying and tracking public opinion and sentiments on particular topics is far from trivial. We have developed a set of tools to interpret and classify opinions, in this paper we concentrate on the use of a lexical resource, a summarization system, and a text analysis tool to create features that enable us to carry out classification of short and long text reviews in the business domain. We apply the techniques to coarse as well as fine grained classification using Support Vector Machines statistical models over data sources in English and Italian. We adopt SVM learning paradigm not only because it has recently been used with success in different tasks in natural language processing, but it has been shown particularly suitable for text categorization (Joachims, 1998). The analytical apparatus and resources are implemented and accessible through the GATE system.

2. Data Sources

For the English language we have collected reviews from three data sources which were appropriate for our application domain (e.,g. business); for the Italian language we are using data provided by the financial news paper Il Sole 24 Ore. The English data sources include (i) a consumer forum where each posting contains a review of a product, service or company with binary rating (DS-I); (ii) a consumer forum where each comment contains a text review and a 5-point rating (DS-II); and (iii) a bank review data source with fine-grained classification (5-point scale) (DS-III). The Italian dataset is a set of over 700 sentence fragments annotated with a fine-grained 5-point scale polarity information. The whole dataset refers to a single Italian company and how its "reputation" evolved in the press (DS-IV). For each data source we carry out a series of experiments transforming each text into a learning instance which is represented as a vector of feature-values, in our case features are created from linguistic annotations produced by different linguistic processors. Each text or sentence has an associated classification allowing us to apply a supervised learning approach to opinion mining.

3. Related Work on Opinion Classification

Classifying product reviews is a common problem in opinion mining and variety of techniques have been used to address the problem including supervised (Li et al., 2007) and unsupervised (Zagibalov and Carroll, 2008) machine-learning. Language resources such as SentiWordNet have recently been developed for the research community (Esuli and Sebastiani, 2006). Some approaches to opinion mining involve predefined gazetteers of positive and negative "opinion words": the well-known Turney's method (Turney, 2002) to determine the semantic orientation of lexemes by calculating their Point-wise Mutual Information (PMI, based on probability of collocations (Church and Hanks, 1990)) to the reference words *excellent* and *poor*. In the financial news domain, (Devitt and Ahmad, 2007) are interested in two problems related to financial news: identifying the polarity of a piece of news, and classifying a text in a fine 7-point scale (from very positive to very negative). They propose a baseline classifier for positive/negative distinction which has an accuracy of 46% and have more so-

phisticated classifiers based on lexical cohesion and SentiWordNet achieving 55% accuracy. (Dave et al., 2003) presents several techniques to create features (words or terms) and associated scores from training corpora for a classification task which consist on sifting positive and negative statements associated to product reviews. Their classifier aggregates features' scores for sentences and bases the classification on the sign of the aggregated score. (Ghose et al., 2007) investigate the issue of generating in an objective way a lexicon of expressions for positive and negative opinion by correlating company gain with reviews.

3.1. SentiWordNet

SentiWordNet (Esuli and Sebastiani, 2006) is a lexical resource in which each synset (set of synonyms) of WordNet (Fellbaum, 1998) is associated with three numerical scores obj (how objective the word is), pos (how positive the word is), and neg (how negative the word is). Each of the scores ranges from 0 to 1, and their sum equals 1. SentiWordNet word values have been semi-automatically computed based on the use of weakly supervised classification algorithms. Examples of "subjectivity" scores associated to WordNet entries are shown in the upper part of Figure 1, the entries contain the parts of speech category of the displayed entry, its positivity, its negativity, and the list of synonyms. We show various synsets related to the words "good" and "bad". There are 4 senses of the noun "good", 21 senses of the adjective "good", and 2 senses of the adverb "good" in WordNet. There is one sense of the noun "bad", 14 senses of the adjective "bad", and 2 senses of the adverb "bad" in WordNet.

In order to identify the positivity or negativity of a given word in text, one first needs to perform word sense disambiguation (WSD). In the current work we ignore WSD and proceed in the following way: for each entry in SentiWordNet (each word#sense or word) we compute the number of times the entry is more positive than negative (positive > negative), the number of times is more negative than positive (positive < negative) and the total number of entries word#sense (or word) in SentiWordNet, therefore we can consider the overall positivity or negativity a particular word has in the lexical resource. We are interested in words that are generally "positive", generally "negative" or generally "neutral" (not much variation between positive and negative). For example a word such as "good" has many more entries where the positive score is greater than the negativity score while a word such as "unhelpful" has more negative occurrences than positive. We use this aggregated scores in our experiments on opinion identification.

We have implemented access to the SentiWordNet resource and provided the above interpretation which allow us to have a "general" sentiment of a word. Access to the resource and the algorithm for text interpretation using the wrapped lexical resource is provided by a plug-in[1].

[1]Note that SentiWordNet requires a licence to be used.

4. Analysis Tools for Feature Computation

In this work, linguistic analysis of textual input is carried out using the General Architecture for Text Engineering (GATE) – a framework for the development and deployment of language processing technology in large scale (Cunningham et al., 2002). For the English data sources we make use of typical GATE components: tokenisation, parts of speech tagging, and morphological analysis. For the Italian data sources we rely on the availability of a parts of speech and lemmatization service that we integrate in our analytical tools in order to produce linguistic annotations for the Italian documents.

After basic analysis of the English documents, we compute the general sentiment of each word based on the interpretation given of SentiWordNet entries. In the bottom part of Figure 1 we show the GATE GUI with a document interpreted with SentiWordNet values. This process will give each word a *positive* and *negative* score, recording also the number of entries the particular word has in SentiWordNet.

In each sentence, the number of words which are more positive than negative (based on the provided scores) is calculated, as it is the number of words which are more negative than positive. These numbers are used to produce a sentiment score for the whole sentence: *positive* if most words in the sentence are positive, *negative* if most words in the sentence are negative, and *neutral* otherwise. At text level, features are created for the number of positive, negative, and neutral sentences in the text or review.

This analysis is complemented with the identification and extraction of adjectives, adverbs, and their bigram combinations which are used as features.

For the Italian case we are only using parts of speech and lemma information (using tf*idf weighting schema) to represent instances, given that we have not implemented access to an Italian lexical resource yet.

5. Experiments and Results

We have carried out standard training and evaluation within a 10-fold cross-validation framework over the four data sets and computed average classification accuracy numbers (sentiment-based features used for the English datasets (DS-I, DS-II, DS-III) and lexical features for the Italian dataset (DS-IV)).

In the binary classification experiments over DS-I we have obtained an average 76% classification accuracy with interesting features emerging such as presence of negative sentiment (as computed by our programs) for the negative class and absence of negative or neutral for the positive class.

In the fine-grained classification experiments over DS-II, we obtained 72% average classification accuracy, here as well with interesting features being picked-up by the classifier, such as absence of negative sentences and

SentiWordNet Fragment

Category	WNT Number	pos	neg	synonyms
a	1006645	0.25	0.375	good#a#15 well#a#2
a	1023448	0.375	0.5	good#a#23 unspoilt#a#1 unspoiled#a#1
a	1073446	0.625	0.0	good#a#22
a	1024262	0.0	1.0	spoilt#a#2 spoiled#a#3 bad#a#4
a	1047353	0.0	0.875	defective#a#3 bad#a#14
a	1074681	0.0	0.875	bad#a#13 forged#a#1

Document Annotated with SentiWordNet Values

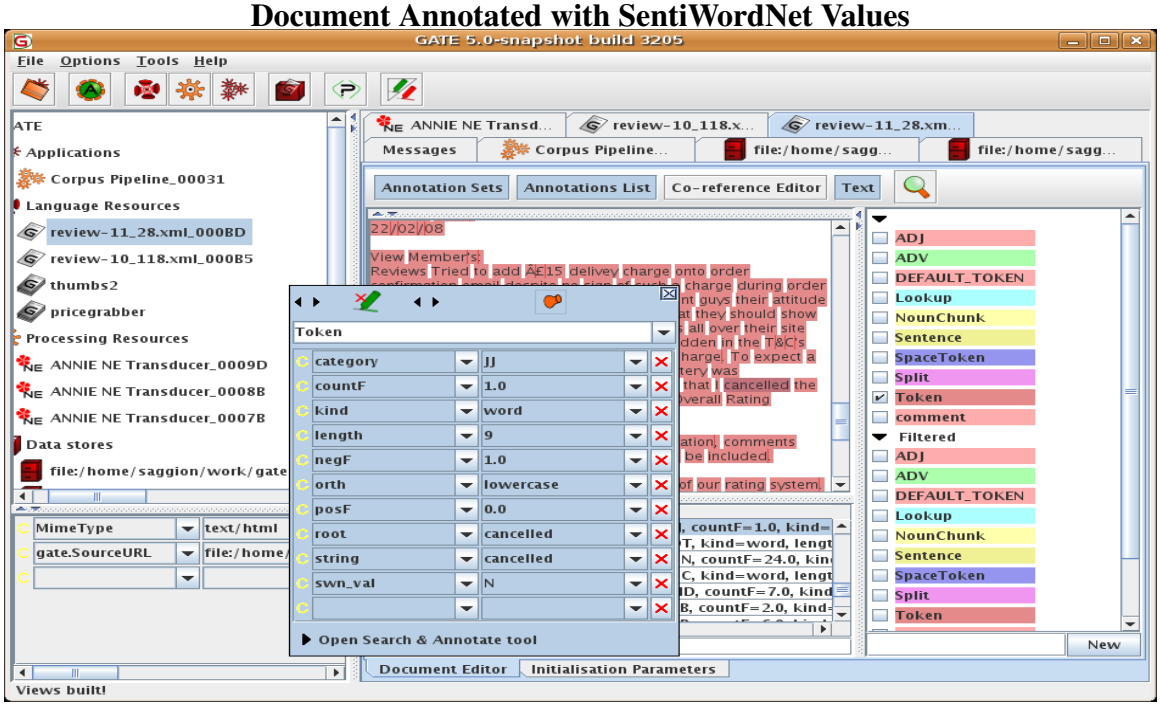

Figure 1: Examples of SentiWordNet Entries and Document Showing SentiWordNet Computed Values.

presence of features such as "very positive" or "happily" for very positive reviews and presence of lexical features such as "cancelled" and "still not" for the very negative reviews.

Results over DS-III (fine-grained classification) were not particularly good, average classification accuracy was 48% which although better than a baseline classification based on the distribution of the most frequent category is far worst than in the other datasets. DS-III texts are much longer than texts in DS-I and DS-II, we have therefore started investigating the use of summaries for reducing the number of features as well as for filtering out noise before classification as explained in Section 5.1.

For the fine-grained classification experiments in Italian (DS-IV) we obtained a maximum classification accuracy of 54% which is worst than the results for English, however, considering the lack of resources for Italian and the fact that the data sources are of a different type (news compared to fora in English) the results can still be considered acceptable specially when compared to past work in the analysis of opinions in news articles.

Overall, results obtained in the four data sets are reasonable and sometimes comparable or even better than previous work in this area, however fair comparison with other approaches is sometimes not possible because of differences in data-sources.

5.1. Summary-based Opinion Classification

Text summarization has been used to support a number of manual and automatic tasks; for example document categorization and question answering (Mani et al., 2002), cross-document coreference (Bagga and Baldwin, 1998; Saggion, 2008a), and semi-automatic essay assessment (Latif and McGee Wood, 2009). Summarization has been studied in the field of sentiment analysis with the objective of producing opinion summaries (TAC, 2008), however, to the best of our knowledge there has been little research on the study of document summarization as a pre-processing step for opinion classification. Given the difficulty of rating reviews on a fine-grained scale using long reviews (DS-III), we decided to summarize the reviews and then use the summaries instead of the full documents to train a classifier. We have experimented with a number of summarization strategies and compression rates to produce summaries (e.g. the percentage of sentences to be extracted from the reviews) and re-

sults of those experiments will be reported elsewhere. Here, we will describe experiments involving query-focused summarization. In order to create summaries for experimentation, we rely on the SUMMA system (Saggion, 2008b), a set of language and processing resources for the creation of summaries which can be used with the GATE system. SUMMA is used to create query-focused summaries of each review in the following way:

- First, we use a corpus statistics module to compute token statistics (e.g. term frequency) relying on a table of inverted frequencies.
- Second, a vector creation module is used to create vectors of terms for each sentence in the review.
- Third, in each review, the name of the entity being reviewed is extracted using a pattern matching procedure; this step extracts strings such as "NatWest", "HSBC Bank", etc.
- Fourth, the strings extracted in the previous step are used as queries to measure sentence relevance in the reviews; the query is transformed in a vector of terms which is compared with each sentence vector in the review using the cosine similarity measure.

This procedure yields a query similarity value for each sentence in the review which in turn is used to rank sentences for an extractive summary. An example of a processed review together with summaries is shown in Figure 2. Summaries at 10%, 20%, 30%, 40%, and 50% compression of the full review were produced by extracting top ranked sentences from the review.

Compression	10%	20%	30%	40%	50%
Accuracy	**56%**	48%	40%	37.5%	35%

Table 1: Classification Accuracy Using Summaries (Accuracy using full documents was 41%).

We then run one experiment per compression rate. Each experiment consisted of predicting the rating of a review using a SVMs trained on word features extracted from the summaries. Various features were used in experiments, here we report numbers corresponding to the use of the following features extracted from the texts: the *root* of each word in the summary, its *category*, and the calculated value employing the *SentiWordNet* lexicon. Here again, we use a 10-fold cross-validation procedure. These features yield a classification accuracy of 41% using the full review. Results of the summarization experiments are presented in Table 1. The best classification result (56%) is obtained with summaries at 10% compression rate. This is better (statistically significant at 90% confidence level) than results obtained using the full document for training the classifier. As the size of the summaries increase classification accuracy decreases for this type of summary. In further experiments we have verified that summaries at 10%, 20% or 30% compression rates provide the best results.

6. Conclusions and Current Work

We have presented a set of tools and experiments for the text-based opinion classification. We have presented a set of GATE tools to compute word-based and sentence-based sentiment features using the SentiWordNet lexical resource. Our experiments show that we can classify short texts in English according to rating (the positive or negative value of the opinions) using machine-learning based on semantic and linguistic analysis. We have also shown experiments for fine-grained polarity classification of sentences in Italian using basic linguistic features obtaining reasonable performance. Classification of long reviews using a 5-point fine-grained scale proved to be a challenging task. We therefore conducted a series of experiments to study the effect that summarization has in sentiment-analysis and more specifically in fine-grained rating of an opinionated text. We have shown that in some cases the use of summaries (e.g. query/entity-focused summaries) could offer competitive advantage over the use of full documents. We are currently working on a new framework for comparison and extrinsic evaluation of summaries in opinion classification, extending the work presented here. In the short term, and for the fine-grained classification task, we intend to replicate the experiment presented here using as evaluation measure "means square errors" which have been pinpointed as a more appropriate measure for classification in an ordinal scale. In the medium to long-term we plan to extent our experiments and analysis to other available datasets in different domains, such as movie or book reviews, in order to see if the results could be influenced by the nature of the corpus.

Acknowledgements

We thank the reviewers of the paper for their comments. We thank Elena Lloret for her collaboration on the summarization experiments. This work was partially supported by the EU-funded MUSING project (IST-2004-027097). The first author is grateful to the Programa Ramón y Cajal 2009 from the Ministerio de Ciencia e Innovación, Spain.

7. References

A. Bagga and B. Baldwin. 1998. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 79–85.

K. W. Church and P. Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

K. Dave, S. Lawrence, and D. M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 519–528, New York, NY, USA. ACM.

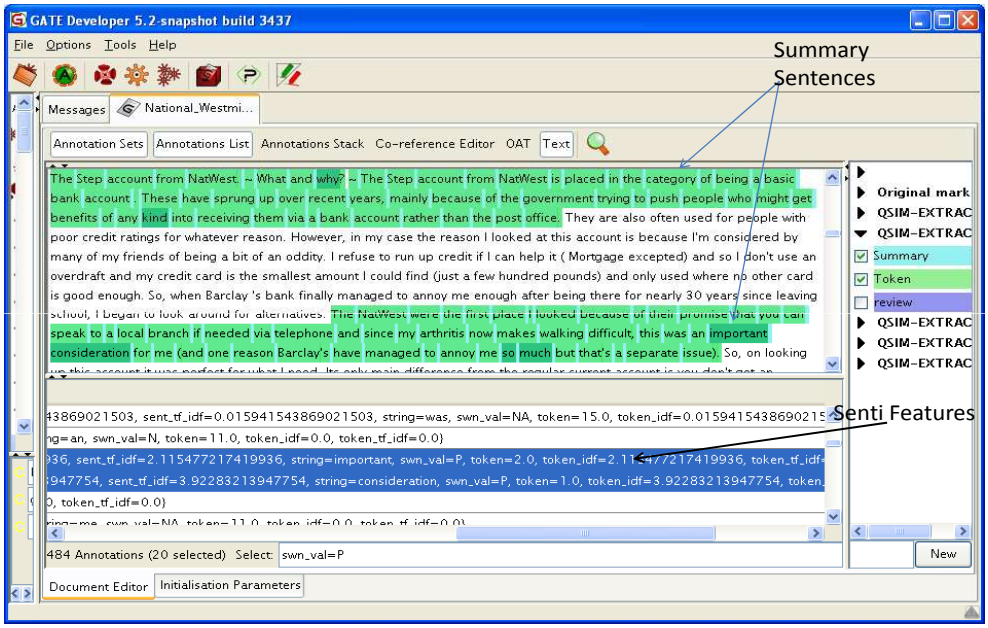

Figure 2: Query-focused summaries, summarization features, and features for classification. Words such as *important* and *consideration* have SentiWordNet value positive.

Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991, Prague, Czech Republic, June. Association for Computational Linguistics.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of LREC-06, 5th Conference on Language Resources and Evaluation*, pages 417–422, Genova, IT.

Christiane Fellbaum, editor. 1998. *WordNet - An Electronic Lexical Database*. MIT Press.

Anindya Ghose, Panagiotis G. Ipeirotis, and Arun Sundararajan. 2007. Opinion mining using econometrics: A case study on reputation systems. In *Proceedings of the Association for Computational Linguistics*. The Association for Computational Linguistics.

T. Joachims. 1998. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398 in Lecture Notes in Computer Science, pages 137–142, Chemnitz, DE. Springer Verlag, Heidelberg, DE.

S. Latif and M. McGee Wood. 2009. A novel technique for automated linguistic quality assessment of students' essays using automatic summarizers. *Computer Science and Information Engineering, World Congress on*, 5:144–148.

Y. Li, K. Bontcheva, and H. Cunningham. 2007. Cost Sensitive Evaluation Measures for F-term Patent Classification. In *The First International Workshop on Evaluating Information Access (EVIA 2007)*, pages 44–53, May.

I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim. 2002. Summac: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68.

H. Saggion and A. Funk. 2009. Extracting opinions and facts for business intelligence. *RNTI*, E-17:119–146.

H Saggion. 2008a. Experiments on Semantic-based Clustering for Cross-document Coreference. In *Proceedings of the 3rd International Join Conference on Natural Language Processing*, Hyderabad, India, January 8-10. AFNLP.

H. Saggion. 2008b. SUMMA: A Robust and Adaptable Summarization Tool. *Traitement Automatique des Langues*, 49(2):103–125.

National Institute of Standards and Technology. 2008. *Proceedings of the Text Analysis Conference*, Gaithesburg, Maryland, USA, November 17-19.

P. D. Turney. 2002. 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)*, pages 417–424, Morristown, NJ, USA, July. Association for Computational Linguistics.

T. Zagibalov and J. Carroll. 2008. Unsupervised classification of sentiment and objectivity in chinese text. In *Proceedings of IJCNLP 2008*, Hyderabad, India, January.