# Experiments on Summary-based Opinion Classification

**Elena Lloret**
Department of Software and Computing Systems
University of Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
elloret@dlsi.ua.es

**Horacio Saggion**
Department of Infomation and Communication Technologies
Grupo TALN
Universitat Pompeu Fabra
C/Tánger, 122-134, 2nd floor
08018 Barcelona, Spain
horacio.saggion@upf.edu

**Manuel Palomar**
Department of Software and Computing Systems
University of Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
mpalomar@dlsi.ua.es

## Abstract

We investigate the effect of text summarisation in the problem of *rating-inference* – the task of associating a fine-grained numerical rating to an opinionated document. We set-up a comparison framework to study the effect of different summarisation algorithms of various compression rates in this task and compare the classification accuracy of summaries and documents for associating documents to classes. We make use of SVM algorithms to associate numerical ratings to opinionated documents. The algorithms are informed by linguistic and sentiment-based features computed from full documents and summaries. Preliminary results show that some types of summaries could be as effective or better as full documents in this problem.

## 1 Introduction

Public opinion has a great impact on company and government decision making. In particular, companies have to constantly monitor public perception of their products, services, and key company representatives to ensure that good reputation is maintained. Recent cases of public figures making headlines for the wrong reasons have shown how companies take into account public opinion to distance themselves from figures which can damage their public image. The Web has become an important source for finding information, in the field of business intelligence, business analysts are turning their eyes to the Web in order to monitor public perception on products, services, policies, and managers. The field of sentiment analysis has recently emerged (Pang and Lee, 2008) as an important area of research in Natural Language Processing (NLP) which can provide viable solutions for monitoring public perception on a number of issues; with evaluation programs such as the *Text REtrieval Conference* track on blog mining [1], the *Text Analysis Conference* [2] track on opinion summarisation, and the *DEfi Fouille de Textes* program (Grouin et al., 2009) advances in the state of the art have been produced. Although sentiment analysis involves various different problems such as identifying subjective sentences or identifying positive and negative opinions in text, here we concentrate on the opinion classification task; and more specifically on *rating-inference*, the task of identifying the author's evaluation of an entity with respect to an ordinal-scale based on the author's textual evaluation of the entity (Pang and Lee, 2005). The specific problem we study in this paper is that of associating a fine-grained rating (1=worst,...5=best) to a review. This is in general considered a difficult problem because of the fuzziness inherent of mid-range ratings (Mukras et al., 2007). A considerable body of research has recently been produced to tackle this problem (Chakraborti et al., 2007; Ferrari et al., 2009) and reported figures showing accuracies ranging from 30% to 50% for such complex task; most approaches derive features for the classification task from the full document. In this research we ask whether extracting features from document summaries could help a classification system. Since text summaries are meant to contain the essential content of a document (Mani, 2001), we investigate whether filtering noise through text summarisation is of any help in the rating-inference task. In re-

[1] http:trec.nist.gov/
[2] http://www.nist.gov/tac/



*Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 107–115,
Los Angeles, California, June 2010. 

cent years, text summarisation has been used to support both manual and automatic tasks; in the SUMMAC evaluation (Mani et al., 1998), text summaries were tested in document classification and question answering tasks where summaries were considered suitable surrogates for full documents; Bagga and Baldwin (1998) studied summarisation in the context of a cross-document coreference task and found that summaries improved the performance of a clustering-based coreference mechanism; more recently Latif and McGee (2009) have proposed text summarisation as a preprocessing step for student essay assessment finding that summaries could be used instead of full essays to group "similar" quality essays. Summarisation has been studied in the field of sentiment analysis with the objective of producing opinion summaries, however, to the best of our knowlegde there has been little research on the study of document summarisation as a text processing step for opinion classification. This paper presents a framework and extensive experiments on text summarisation for opinion classification, and in particular, for the rating-inference problem. We will present results indicating that some types of summaries could be as effective or better than the full documents in this task.

The remainder of the paper is organised as follows: Section 2 will compile the existing work with respect to the inference-rating problem; Section 3 and Section 4 will describe the corpus and the NLP tools used for all the experimental set-up. Next, the text summarisation approaches will be described in Section 5, and then Section 6 will show the experiments conducted and the results obtained together with a discussion. Finally, we will draw some conclusions and address further work in Section 7.

## 2 Related Work

Most of the literature regarding sentiment analysis addresses the problem either by detecting and classifying opinions at a sentence level (Wilson et al., 2005; Du and Tan, 2009), or by attempting to capture the overall sentiment of a document (McDonald et al., 2007; Hu et al., 2008). Traditional approaches tackle the task as binary classification, where text units (e.g. words, sentences, fragments) are classified into *positive vs. negative*, or *subjective vs. objective*, according to their polarity and subjectivity degree, respectively. However, sentiment classification taking into account a finer granularity has been less considered. Rating-inference is a particular task within sentiment analysis, which aims at inferring the author's numerical rating for a review. For instance, given a review and 5-star-rating scale (ranging from 1 -the worst- to 5 -the best), this task should correctly predict the review's rating, based on the language and sentiment expressed in its content.

In (Pang and Lee, 2005), the rating-inference problem is analysed for the movies domain. In particular, the utility of employing label and item similarity is shown by analysing the performance of three different methods based on SVM (one vs. all, regression and metric labeling), in order to infer the author's implied numerical rating, which ranges from 1 up to 4 stars, depending on the degree the author of the review liked or not the film. The approach described in (Leung et al., 2006) suggests the use of collaborative filtering algorithms together with sentiment analysis techniques to obtain user preferences expressed in textual reviews, focusing also on movie reviews. Once the opinion words from user reviews have been identified, the polarity of those opinion words together with their strength need to be computed and mapped to the rating scales to be further input to the collaborative input algorithms.

Apart from these approaches, this problem is stated from a different point of view in (Shimada and Endo, 2008). Here it is approached from the perspective of rating different details of a product under the same review. Consequently, they rename the problem as "*seeing several stars*" instead of only one, corresponding to the overall sentiment of the review. Also, in (Baccianella et al., 2009) the rating of different features regarding hotel reviews (cleanliness, location, staff, etc.) is addressed by analysing several aspects involved in the generation of product review's representations, such as part-of-speech and lexicons. Other approaches (Devitt and Ahmad, 2007), (Turney, 2002) face this problem by grouping documents with closer stars under the same category, i.e. positive or negative, simplifying the task into a binary classification problem.

Recently, due to the vast amount of on-line information and the subjectivity appearing in documents, the combination of sentiment analysis and summari-

sation task in tandem can result in great benefits for stand-alone applications of sentiment analysis, as well as for the potential uses of sentiment analysis as part of other NLP applications (Stoyanov and Cardie, 2006). Whilst there is much literature combining sentiment analysis and text summarisation focusing on generating opinion-oriented summaries for the new textual genres, such as blogs (Lloret et al., 2009), or reviews (Zhuang et al., 2006), the use of summaries as substitutes of full documents in tasks such as rating-inference has been not yet explored to the best of our knowledge. In contrast to the existing literature, this paper uses summaries instead of full reviews to tackle the rating-inference task in the financial domain, and we carry out a preliminary analysis concerning the potential benefits of text summaries for this task.

## 3 Dataset for the Rating-inference Task

Since there is no standard dataset for carrying out the rating-inference task, the corpus used for our experiments was one associated to a current project on business intelligence we are working on. These data consisted of 89 reviews of several English banks (Abbey, Barcalys, Halifax, HSBC, Lloyds TSB, and National Westminster) gathered from the Internet. In particular the documents were collected from *Ciao*[3], a Website where users can write reviews about different products and services, depending on their own experience.

Table 1 lists some of the statistical properties of the data. It is worth stressing upon the fact that the reviews have on average 2,603 words, which means that we are dealing with long documents rather than short ones, making the rating-inference task even more challenging. The shortest document contains 1,491 words, whereas the longest document has more than 5,000 words.

| # Reviews | Avg length | Max length | Min length |
|---|---|---|---|
| 89 | 2,603 | 5,730 | 1,491 |

Table 1: Corpus Statistics

Since the aim of the task we are pursuing focuses on classifying correctly the star for a review (ranging from 1 to 5 stars), it is necessary to study how many reviews we have for each class, in order to see whether we have a balanced distribution or not. Table 2 shows this numbers for each star-rating. It is worth mentioning that one-third of the reviews belong to the 4-star class. In contrast, we have only 9 reviews that have been rated as 3-star, consisting of the 10% of the corpus, which is a very low number.

| Star-rating | # reviews | % |
|---|---|---|
| 1-star | 17 | 19 |
| 2-star | 11 | 12 |
| 3-star | 9 | 10 |
| 4-star | 28 | 32 |
| 5-star | 24 | 27 |

Table 2: Class Distribution

## 4 Natural Language Processing Tools

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). We make use of typical GATE components: tokenisation, parts of speech tagging, and morphological analysis to produce document annotations. From the annotations we produce a number of features for document representation. Features produced from the annotations are: *string* – the original, unmodified text of each token; *root* – the lemmatised, lower-case form of the token; *category* – the part-of-speech (POS) tag, a symbol that represents a grammatical category such as determiner, present-tense verb, past-tense verb, singular noun, etc.; *orth* – a code representing the token's combination of upper- and lower-case letters. In addition to these basic features, "sentiment" features based on a lexical resource are computed as explained below.

### 4.1 Sentiment Features

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 classi-

[3] http://www.ciao.co.uk/

fication algorithms. In this work we compute the "general sentiment" of a word in the following way: given a word $w$ we compute the number of times the word $w$ is more positive than negative (positive $>$ negative), the number of times is more negative than positive (positive $<$ negative) and the total number of entries of word $w$ in SentiWordNet, therefore we can consider the overall positivity or negativity a particular word has in SentiWordNet. 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 classification experiments. Note that we do not apply any word sense disambiguation procedure here.

### 4.2 Machine Learning Tool

For the experiments reported here, we adopt a Support Vector Machine (SVM) learning paradigm not only because it has recently been used with success in different tasks in natural language processing (Isozaki and Kazawa, 2002), but it has been shown particularly suitable for text categorization (Kumar and Gopal, 2009) where the feature space is huge, as it is in our case. We rely on the support vector machines implementation distributed with the GATE system (Li et al., 2009) which hides from the user the complexities of feature extraction and conversion from documents to the machine learning implementation. The tool has been applied with success to a number of datasets for opinion classification and rating-inference (Saggion and Funk, 2009).

## 5 Text Summarisation Approach

In this Section, three approaches for carrying out the summarisation process are explained in detail. First, a generic approach is taken as a basis, and then, it is adapted into a query-focused and a opinion-oriented approach, respectively.

### 5.1 Generic Summarisation

A generic text summarisation approach is first taken as a core, in which three main stages can be distinguished: i) document preprocessing; ii) relevance detection; and ii) summary generation. Since we work with Web documents, an initial preprocessing step is essential to remove all unnecessary tags and noisy information. Therefore, in the first stage the body of the review out of the whole Web page is automatically delimitated by means of patterns, and only this text is used as the input for the next summarisation stages. Further on, a sentence relevance detection process is carried out employing different combinations of various techniques. In particular, the techniques employed are:

**Term frequency (*tf*)**: this technique has been widely used in different summarisation approaches, showing the the most frequent words in a document contain relevant information and can be indicative of the document's topic (Nenkova et al., 2006)

**Textual entailment (*te*)**: a *te* module (Ferrández et al., 2007) is used to detect redundant information in the document, by computing the entailment between two consecutive sentences and discarding the entailed ones. The identification of these entailment relations helps to avoid incorporating redundant information in summaries.

**Code quantity principle (*cqp*)**: this is a linguistic principle which proves the existence of a proportional relation between how important the information is, and the number of coding elements it has (Givón, 1990). In this approach we assume that sentences containing longer noun-phrases are more relevant.

The aforementioned techniques are combined together taking always into account the term-frequency, leading to different summarisation strategies (*tf*, *te+tf*, *cqp+tf*, *te+cqp+tf*). Finally, the resulting summary is produced by extracting the highest scored sentences up to the desired length, according the techniques explained.

### 5.2 Query-focused Summarisation

Through adapting the generic summarisation approach into a query-focused one, we could benefit from obtaining more specific sentences with regard to the topic of the review. As a preliminary work, we are going to assume that a review is about a bank, and as a consequence, the name of the bank is considered to be the topic. It is worth mentioning that a person can refer to a specific bank in different ways. For example, in the case of "*The National Westmin-*

*ster Bank"*, it can be referred to as *"National Westminster"* or *"NatWest"*. Such different denominations were manually identified and they were used to biased the content of the generated summaries, employing the same techniques of *tf*, *te* and the *cqp* combined together. One limitation of this approach is that we do not directly deal with the coreference problem, so for example, sentences containing pronouns referring also to the bank, will not be taken into consideration in the summarisation process. We are aware of this limitation and for future work it would be necessary to run a coreference algorithm to identify all occurrences of a bank within a review. However, since the main goal of this paper is to carry out a preliminary analysis of the usefulness of summaries in contrast to whole reviews in the rating-inference problem, we did not take this problem into account at this stage of the research. In addition, when we do query-focused summarisation only we rely on the SUMMA toolkit (Saggion, 2008) to produce a query similarity value for each sentence in the review which in turn is used to rank sentences for an extractive summary (*qf*). This similarity value is the cosine similarity between a sentence vector (terms and weights) and a query vector (terms and weigths) and where the query is the name of the entity being reviewed (e.g. *National Westminster*).

### 5.3 Opinion-oriented Summarisation

Since reviews are written by people who want to express their opinion and experience with regard to a bank, in this particular case, either generic or query-focused summaries can miss including some important information concerning their sentiments and feelings towards this particular entity. Therefore, a sentiment classification system similar to the one used in (Balahur-Dobrescu et al., 2009) is used together with the summarisation approach, in order to generate opinion-oriented summaries. First of all, the sentences containing opinions are identified, assigning each of them a polarity (positive and negative) and a numerical value corresponding to the polarity strength (the higher the negative score, the more negative the sentence and similarly, the higher the positive score, the more positive the sentence). Sentences containing a polarity value of 0 are considered neutral and are not taken into account. Once the sentences are classified into positives, negatives and neutrals, they are grouped together according to its type. Further on, the same combination of techniques as for previously explained summarisation approaches are then used.

Additionally, a summary containing only the most positive and negative sentences is also generated (we have called this type of summaries *sent*) in order to check whether the polarity strength on its own could be a relevant feature for the summarisation process.

## 6 Evaluation Environment

In this Section we are going to describe in detail all the experimental set-up. Firstly, we will explain the corpus we used together with some figures regarding some statistics computed. Secondly, we will describe in-depth all the experiments we ran and the results obtained. Finally, an extensive discussion will be given in order to analyse all the results and draw some conclusions.

### 6.1 Experiments and Results

The main objective of the paper is to investigate the influence of summaries in contrast to full reviews for the rating-inference problem.

The purpose of the experiments is to analyse the performance of the different suggested text summarisation approaches and compare them to the performance of the full review. Therefore, the experiments conducted were the following: for each proposed summarisation approach, we experimented with five different types of compression rates for summaries (ranging from 10% to 50%). Apart from the full review, we dealt with 14 different summarisation approaches (4 for generic, 5 for query-focused and 5 for opinion-oriented summarisation), as well as 2 baselines (*lead* and *final*, taking the first or the last sentences according to a specific compression rate, respectively). Each experiment consisted of predicting the correct star of a review, either with the review as a whole or with one of the summarisation approaches. As we previously said in Section 4, for predicting the correct star-rating, we used machine learning techniques. In particular, different features were used to train a SVM classifier with 10-fold cross validation[4], using the whole review:

[4]The classifier used was the one integrated within the GATE framework: http://gate.ac.uk/

the *root* of each word, its *category*, and the calculated value employing the *SentiWordNet* lexicon, as well as their combinations. As a baseline for the full document we took into account a totally uninformed approach with respect to the class with higher number of reviews, i.e. considering all documents as if they were scored with 4 stars. The different results according different features can be seen in Table 3.

| Feature | $F_{\beta=1}$ |
|---|---|
| *baseline* | 0.300 |
| *root* | 0.378 |
| *category* | 0.367 |
| *sentiWN* | 0.333 |
| *root+category* | 0.356 |
| *root+sentiWN* | 0.333 |
| *category+sentiWN* | 0.389 |
| *root+category+sentiWN* | **0.413** |

Table 3: F-measure results using the full review for classification

Regarding the features for training the summaries, it is worth mentioning that the best performing feature when no sentiment-based features are taken into account is the one using the root of the words. Consequently, this feature was used to train the summaries. Moreover, since the best results using the full review were obtained using the combination of the all the features (*root+category+sentiWN*), we also selected this combination to train the SVM classifier with our summaries. Conducting both experiments, we could analyse to what extent the sentiment-based feature benefit the classification process.

The results obtained are shown in Table 4 and Table 5, respectively. These tables show the F-measure value obtained for the classification task, when features extracted from summaries are used instead from the full review. On the one hand, results using the *root* feature extracted from summaries can be seen in Table 4. On the other hand, Table 5 shows the results when the combination of all the linguistic and sentiment-based features (*root+category+sentiWN*), that has been extracted from summaries, are used for training the SVM classifier.

We also performed two statistical tests in order to measure the significance for the results obtained. The tests we performed were the one-way Analysis of Variance (ANOVA) and the t-test (Spiegel and Castellan, 1998). Given a group of experiments, we first run ANOVA for analysing the difference between their means. In case some differences are found, we run the t-test between those pairs.

### 6.2 Discussion

A first analysis derived from the results obtained in Table 3 makes us be aware of the difficulty associated to the rating-inference task. As can be seen, a baseline without any information from the document at all, is performing around 30%, which compared to the remaining approaches is not a very bad number. However, we assumed that dealing with some information contained in documents, the classification algorithm will do better in finding the correct star associated to a review. This was the reason why we experimented with different features alone or in combination. From these experiments, we obtained that the combination of linguistic and semantic-based features leads to the best results, obtaining a F-measure value of 41%. If sentiment-based features are not taken into account, the best feature is the root of the word on its own. Furthermore, in order to analyse further combinations, we ran some experiments with bigrams. However, the results obtained did not improve the ones we already had, so they are not reported in this paper.

As far as the results is concerned comparing the use of summaries to the full document, it is worth mentioning that when using specific summarisation approaches, such as query-focused summaries combined with term-frequency, we get better results than using the full document with a 90% confidence interval, according to a t-test. In particular, *qf* for 10% is significant with respect to the full document, using only root as feature for training. For the results regarding the combination of *root*, *category* and *SentiWordNet*, *qf* for 10% and *qf+tf* for 10% and 20% are significant with respect to the full document.

Concerning the different summarisation approaches, it cannot be claimed a general tendency about which ones may lead to the best results. We also performed some significance tests between different strategies, and in most of the cases, the t-test and the ANOVA did not report significance over 95%. Only a few approaches were significant at a 95% confidence level, for instance, *te+cqp+tf* and *sent+te+cqp+tf* with respect to *sent+cqp+tf*

| Approach | | Compression Rate | | | | |
|---|---|---|---|---|---|---|
| **Summarisation method** | | **10%** | **20%** | **30%** | **40%** | **50%** |
| lead | $F_{\beta=1}$ | 0.411 | 0.378 | 0.367 | 0.311 | 0.322 |
| final | $F_{\beta=1}$ | 0.322 | 0.389 | 0.300 | **0.467** | **0.456** |
| tf | $F_{\beta=1}$ | 0.400 | 0.344 | 0.400 | 0.367 | 0.367 |
| te+tf | $F_{\beta=1}$ | 0.367 | 0.422 | 0.411 | 0.389 | 0.322 |
| cqp+tf | $F_{\beta=1}$ | 0.300 | 0.344 | 0.311 | 0.300 | 0.256 |
| te+cqp+tf | $F_{\beta=1}$ | 0.422 | 0.356 | 0.333 | 0.300 | 0.322 |
| qf | $F_{\beta=1}$ | 0.513 | 0.388 | 0.375 | 0.363 | 0.363 |
| qf+tf | $F_{\beta=1}$ | **0.567** | **0.467** | 0.311 | 0.367 | 0.389 |
| qf+te+tf | $F_{\beta=1}$ | 0.389 | 0.367 | 0.411 | 0.378 | 0.333 |
| qf+cqp+tf | $F_{\beta=1}$ | 0.300 | 0.356 | 0.378 | 0.378 | 0.333 |
| qf+te+cqp+tf | $F_{\beta=1}$ | 0.322 | 0.322 | 0.367 | 0.367 | 0.356 |
| sent | $F_{\beta=1}$ | 0.344 | 0.380 | 0.391 | 0.290 | 0.336 |
| sent+tf | $F_{\beta=1}$ | 0.378 | 0.425 | **0.446** | 0.303 | 0.337 |
| sent+te+tf | $F_{\beta=1}$ | 0.278 | 0.424 | 0.313 | 0.369 | 0.347 |
| sent+cqp+tf | $F_{\beta=1}$ | 0.333 | 0.300 | 0.358 | 0.358 | 0.324 |
| sent+te+cqp+tf | $F_{\beta=1}$ | 0.446 | 0.334 | 0.358 | 0.292 | 0.369 |

Table 4: Classification results (F-measure) for summaries using *root* (*lead* = first sentences; *final* = last sentences; *tf* = term frequency; *te* = textual entailment; *cqp* = code quantity principle with noun-phrases; *qf* = query-focused summaries; and *sent* = opinion-oriented summaries)

for 10%; *sent+tf* in comparison to *sent+cqp+tf* for 20%; or *sent* with respect to *cqp+tf* for 40% and 50% compression rates. Other examples of the approaches that were significant at a 90% level of confidence are *qf* for 10% with respect to *sent+te+cqp+tf*. Due to the wide range of summarisation strategies tested in the experiments, the results obtained vary a lot and, due to the space limitations, it is not possible to report all the tables. What it seems to be clear from the results is that the code quantity principle (see Section 5) is not contributing much to the summarisation process, thus obtaining poor results when it is employed. Intuitively, this can be due to the fact that after the first mention of the bank, there is a predominant use of pronouns, and as a consequence, the accuracy of the tool that identifies noun-phrases could be affected. The same reason could be affecting the term-frequency calculus, as it is computed based on the lemmas of the words, not taking into account the pronouns that refer also to them.

## 7 Conclusion and Future Work

This paper presented a preliminary study of inference-rating task. We have proposed here a new framework for comparison and extrinsic evaluation of summaries in a text-based classification task. In our research, text summaries generated using different strategies were used for training a SVM classifier instead of full reviews. The aim of this task was to correctly predict the category of a review within a 1 to 5 star-scale. For the experiments, we gathered 89 bank reviews from the Internet and we generated 16 summaries of 5 different compression rates for each of them (80 different summaries for each review, having generated in total 7,120 summaries). We also experimented with several linguistic and sentiment-based features for the classifier. Although the results obtained are not significant enough to state that summaries really help the rating-inference task, 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 and we have also shown that some summarisation techniques do not degrade the performance of a rating-inference algorithm when compared to the use of full documents. We strongly believe that this preliminary study could serve as a starting point for future developments.

Although we have carried out extensive experimentation with different summarisation techniques, compression rates, and document/summary features, there are many issues that we have not explored. In the future, we plan to investigate whether the results could be affected by the class distribution of the reviews, and in this line we would like to see the distribution of the documents using clustering tech-

| Approach | | Compression Rate | | | | |
|---|---|---|---|---|---|---|
| **Summarisation method** | | **10%** | **20%** | **30%** | **40%** | **50%** |
| lead | $F_{\beta=1}$ | 0.275 | 0.422 | 0.422 | 0.378 | 0.322 |
| final | $F_{\beta=1}$ | 0.275 | 0.378 | 0.333 | 0.344 | 0.400 |
| tf | $F_{\beta=1}$ | 0.411 | 0.422 | 0.411 | 0.378 | 0.378 |
| te+tf | $F_{\beta=1}$ | 0.411 | 0.344 | 0.344 | 0.344 | 0.378 |
| cqp+tf | $F_{\beta=1}$ | 0.358 | 0.267 | 0.333 | 0.222 | 0.289 |
| te+cqp+tf | $F_{\beta=1}$ | 0.444 | 0.411 | 0.411 | 0.311 | 0.322 |
| qf | $F_{\beta=1}$ | **0.563** | **0.488** | 0.400 | 0.375 | 0.350 |
| qf+tf | $F_{\beta=1}$ | 0.444 | 0.411 | **0.433** | 0.367 | 0.356 |
| qf+te+tf | $F_{\beta=1}$ | 0.322 | 0.367 | 0.356 | 0.344 | 0.344 |
| qf+cqp+tf | $F_{\beta=1}$ | 0.292 | 0.322 | 0.367 | 0.333 | 0.356 |
| qf+te+cqp+tf | $F_{\beta=1}$ | 0.356 | 0.378 | 0.356 | 0.367 | 0.356 |
| sent | $F_{\beta=1}$ | 0.322 | 0.370 | 0.379 | **0.412** | **0.414** |
| sent+tf | $F_{\beta=1}$ | 0.378 | 0.446 | 0.359 | 0.380 | 0.402 |
| sent+te+tf | $F_{\beta=1}$ | 0.333 | 0.414 | 0.404 | 0.380 | 0.381 |
| sent+cqp+tf | $F_{\beta=1}$ | 0.300 | 0.333 | 0.347 | 0.358 | 0.296 |
| sent+te+cqp+tf | $F_{\beta=1}$ | 0.436 | 0.413 | 0.425 | 0.359 | 0.324 |

Table 5: Classification results (F-measure) for summaries using *root*, *category* and *SentiWordNet* (*lead* = first sentences; *final* = last sentences; *tf* = term frequency; *te* = textual entailment; *cqp* = code quantity principle with noun-phrases; *qf* = query-focused summaries; and *sent* = opinion-oriented summaries)

niques. Moreover, we would also like to investigate what it would happen if we consider the values of the star-rating scale as ordinal numbers, and not only as labels for categories. We will replicate the experiments presented here using as evaluation measure the "mean square error" which has been pinpointed as a more appropriate measure for categorisation in an ordinal scale. Finally, in the medium to long-term we plan to extent the 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, allowing also further results for comparison with other approaches and assessing the difficulty of the task from a perspective of different domains.

## Acknowledgments

This research has been supported by the project PROMETEO "Desarrollo de Técnicas Inteligentes e Interactivas de Minería de Textos" (2009/119) from the Valencian Government. Moreover, Elena Lloret is funded by the FPI program (BES-2007-16268) from the Spanish Ministry of Science and Innovation under the project TEXTMESS (TIN2006-15265-C06-01), and Horacio Saggion is supported by a Ramón y Cajal Fellowship from the Ministry of Science and Innovation, Spain. The authors would also like to thank Alexandra Balahur for helping to process the dataset with her Opinion Mining approach.

## References

S. Baccianella, A. Esuli, and F. Sebastiani. 2009. Multi-facet Rating of Product Reviews. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 461–472.

A. Bagga and B. Baldwin. 1998. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proceedings of the COLING-ACL*, pages 79–85.

A. Balahur-Dobrescu, M. Kabadjov, J. Steinberger, R. Steinberger, and A. Montoyo. 2009. Summarizing Opinions in Blog Threads. In *Proceedings of the Pacific Asia Conference on Language, INformation and Computation Conference*, pages 606–613.

S. Chakraborti, R. Mukras, R. Lothian, N. Wiratunga, S. Watt, and D Harper. 2007. Supervised Latent Semantic Indexing using Adaptive Sprinkling. In *Proceedings of IJCAI-07*, pages 1582–1587.

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 ACL*.

A. Devitt and K. Ahmad. 2007. Sentiment Polarity Identification in Financial News: A Cohesion-based Approach. In *Proceedings of the ACL*, pages 984–991.

W. Du and S. Tan. 2009. An Iterative Reinforcement Approach for Fine-Grained Opinion Mining. In *Proceedings of the NAACL*, pages 486–493.

A. Esuli and F. Sebastiani. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of LREC*, pages 417–422.

C. Fellbaum. 1998. *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA.

O. Ferrández, D. Micol, R. Muñoz, and M. Palomar. 2007. A Perspective-Based Approach for Solving Textual Entailment Recognition. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 66–71, June.

S. Ferrari, T. Charnois, Y. Mathet, F. Rioult, and D. Legallois. 2009. Analyse de Discours Évaluatif, Modèle Linguistique et Applications. In *Fouille de données d'opinion*, volume E-17, pages 71–93.

T. Givón, 1990. *Syntax: A functional-typological introduction, II*. John Benjamins.

C. Grouin, M. Hurault-Plantet, P. Paroubek, and J. B. Berthelin. 2009. DEFT'07 : Une Campagne d'Avaluation en Fouille d'Opinion. In *Fouille de données d'opinion*, volume E-17, pages 1–24.

Y. Hu, W. Li, and Q. Lu. 2008. Developing Evaluation Model of Topical Term for Document-Level Sentiment Classification. In *Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence*, pages 175–186.

H. Isozaki and H. Kazawa. 2002. Efficient Support Vector Classifiers for Named Entity Recognition. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 390–396.

M. A. Kumar and M. Gopal. 2009. Text Categorization Using Fuzzy Proximal SVM and Distributional Clustering of Words. In *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 52–61.

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.

C. W. K. Leung, S. C. F. Chan, and F. L. Chung. 2006. Integrating Collaborative Filtering and Sentiment Analysis: A Rating Inference Approach. In *Proceedings of The ECAI 2006 Workshop on Recommender Systems*, pages 62–66.

Y. Li, K. Bontcheva, and H. Cunningham. 2009. Adapting SVM for Data Sparseness and Imbalance: A Case Study in Information Extraction. *Natural Language Engineering*, 15(2):241–271.

E. Lloret, A. Balahur, M. Palomar, and A. Montoyo. 2009. Towards Building a Competitive Opinion Summarization System: Challenges and Keys. In *Proceedings of the NAACL. Student Research Workshop and Doctoral Consortium*, pages 72–77.

I. Mani, D. House, G. Klein, L. Hirshman, L. Obrst, T. Firmin, M. Chrzanowski, and B. Sundheim. 1998. The TIPSTER SUMMAC Text Summarization Evaluation. Technical report, The Mitre Corporation.

I. Mani. 2001. *Automatic Text Summarization*. John Benjamins Publishing Company.

R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar. 2007. Structured Models for Fine-to-Coarse Sentiment Analysis. In *Proceedings of the ACL*, pages 432–439.

R. Mukras, N. Wiratunga, R. Lothian, S. Chakraborti, and D. Harper. 2007. Information Gain Feature Selection for Ordinal Text Classification using Probability Redistribution. In *Proceedings of the Textlink workshop at IJCAI-07*.

A. Nenkova, L. Vanderwende, and K. McKeown. 2006. A Compositional Context Sensitive Multi-document Summarizer: Exploring the Factors that Influence Summarization. In *Proceedings of the ACM SIGIR conference on Research and development in information retrieval*, pages 573–580.

B. Pang and L. Lee. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the ACL*, pages 115–124.

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

H. Saggion and A. Funk. 2009. Extracting Opinions and Facts for Business Intelligence. *RNTI*, E-17:119–146.

H. Saggion. 2008. SUMMA: A Robust and Adaptable Summarization Tool. *Traitement Automatique des Languages*, 49:103–125.

K. Shimada and T. Endo. 2008. Seeing Several Stars: A Rating Inference Task for a Document Containing Several Evaluation Criteria. In *Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 1006–1014.

S. Spiegel and N. J. Castellan, Jr. 1998. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill International.

V. Stoyanov and C. Cardie. 2006. Toward Opinion Summarization: Linking the Sources. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 9–14.

P. D. Turney. 2002. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the ACL*, pages 417–424.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the EMNLP*, pages 347–354.

L. Zhuang, F. Jing, and X. Y. Zhu. 2006. Movie Review Mining and Summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50.