Investigating the Use of Extractive Summarisation in Sentiment Classification

Marco Bonzanini, Miguel Martinez-Alvarez, and Thomas Roelleke

Queen Mary, University of London
Mile End Road, E1 4NS London, UK
{marcob,miguel,thor}@eecs.qmul.ac.uk

Abstract. In online reviews, authors often use a short passage to describe the overall feeling about a product or a service. A review as a whole can mention many details not in line with the overall feeling, so capturing this key passage is important to understand the overall sentiment of the review. This paper investigates the use of extractive summarisation in the context of sentiment classification. The aim is to find the summary sentence, or the short passage, which gives the overall sentiment of the review, filtering out potential noisy information. Experiments are carried out on a movie review data-set. The main finding is that subjectivity detection plays a central role in building summaries for sentiment classification. Subjective extracts carry the same polarity of the full text reviews, while statistical and positional approaches are not able to capture this aspect.

1 Introduction

The popularity of on-line resources, which allow users to review products or services, is motivating new interest in the area of Sentiment Analysis [10]. One of the main tasks in this field is the classification of opinionated documents according to the overall sentiment, i.e. whether positive or negative. A common behaviour among reviewers is to summarise the overall sentiment of the review in a single sentence, or in a short passage. On the other hand, the rest of the review can express a feeling which is different from the overall judgement. This can be explained by the presence of several aspects or features that the reviewers want to comment on. As an example, we can consider the following review, taken from RottenTomatoes[1], a popular movie review site. The words or phrases carrying opinions are marked in italic. Several sentences express disappointment about different aspects of the movie, and simply counting the negative sentences would lead to classify the review as negative. The overall recommendation, described in the last sentence, is instead positive. It is also worth noting that some expressions, like "too easily", do not carry a negative sentiment per se, but must be put into context to be understood. In a similar way, terms normally related to negative feelings, like "trauma", are not used to denote a negative opinion:

[1] http://www.rottentomatoes.com

> I was particularly *disappointed* that the film didn't deal more with the trauma of learning one's life is a tv show [...] I almost felt that he got over it *too easily* for the sake of the film's pacing [...] Perhaps it's not fair to criticize a movie for what it isn't, but it seems like there were *some missed opportunities* here. But on its own terms, the movie is *well made*.

Moreover, often a review contains sentences which do not provide any information about opinions, i.e. they are not subjective. This is the case of movie reviews, where a short picture of the plot can be given to open the review, without commenting on it. Previous work has shown how the capability of identifying subjective sentences can improve the sentiment classification [9].

This paper investigates how the use of summarisation techniques can be applied in the context of sentiment classification of on-line reviews. The focus is on the movie review domain, which is considered to be particularly challenging, as people write not only about the movie itself, but also about movie elements such as special effects or music, and about movie-related people [15]. More specifically, the aim is to capture the summary passage, i.e. the short passage, or even the single sentence, which gives the overall sentiment of the review. From the user's perspective, the advantage of having a summarised review consists in a reduced effort to understand the message of the document, given that the key information is preserved. Traditional sentence extraction techniques can be applied for this task, although a more opinion-oriented approach is needed, since the goal is not to better describe the topic of the review in a single sentence, but to capture its overall polarity. In order to verify whether the summarisation task preserves the information about the sentiment of reviews, text classification is performed on the original documents and on the produced summaries.

The contributions of this work are two-fold: firstly, we show how the summaries based on subjectivity well represent the polarity of the full-text review; secondly, we investigate different techniques for identifying the key passage of a review with respect to polarity. Experiments on a movie review data-set show the importance of subjectivity detection for polarity classification.

The rest of the paper is organised as follows. Section 2 presents the related work on sentiment summarisation, and classification through summarisation. In Section 3 the overall approach for sentiment classification and summarisation is proposed. Section 4 reports the experimental study, and Section 5 concludes the paper outlining the directions for future work.

2 Related Work

Previous work in summarisation of opinionated documents has been focusing on different domains of user-generated content. Dealing with short web comments, an approach for extracting the top sentiment keywords and for showing them in a tag cloud, has been proposed in [11]. This approach is based on the use of Pointwise Mutual Information (PMI) as described in [14]. Experiments in the context of digital product reviews have

been reported in [4]. This technique uses a set of seed adjectives of known polarity, which is expanded with the use of WordNet (i.e. synonyms share the same polarity, while antonyms have the opposite polarity). The generation of summaries consists then in aggregating opinionated sentences related to the same feature. A multi-knowledge approach has been shown in [15], with experiments on the movie review domain. This approach aims at identifying movie features, like the soundtrack or the photography, as well as movie-related people, like actors, director, etc. Since single opinions can be expressed on a specific feature of a movie, their approach can be used to build personalised feature-oriented summaries. A similar work has been proposed in [2] in the context of local service reviews.

The use of summarisation to improve classification has been explored in [13]. Different summarisation techniques can be applied to generate summaries of web-page, resulting in an improvement of their classification. This approach differs from the one proposed in this paper, as they face the problem of topic classification rather than sentiment classification. The work presented in [3] implies the use of sentence extraction techniques, although it is not focused on summarisation per se. The use of sentence-level evidences, in particular the location of the sentence within the document, is used to improve opinion retrieval. In this approach, relevance and polarity are combined to retrieve blog posts.

The idea of a single sentence extraction, to determine the polarity of the whole document, has been suggested in [1], although results on the polarity classification task have not been reported. Another summarisation approach, based on subjectivity detection, is shown in [9]. The main idea is to filter out the objective sentences, i.e. the ones not carrying sentiment information, and to base the polarity classification entirely on the subjective sentences. Proximity information is also taken into account, as subjective sentences tend to be close to each other. This method has been shown to significantly improve the classification, compared to the results of a Naive Bayes classifier on the whole document, and to be not significantly worse than a Support Vector Machine classifier.

3 Methodology

This section describes the main components of the proposed approach, namely a sentiment classifier, an extractive summariser and a subjectivity classifier. Figure 1 describes the pipeline for the movie review classification. The reviews can be classified directly (full text) or can be summarised in three different ways. Firstly, through the summarisation component, sentence extraction based on statistical or positional approaches is performed. Secondly, through the subjectivity detection component, objective sentences are filtered out, keeping all and only the subjective ones to form the summary. Thirdly, through a pipeline of both components, subjective extracts are further summarised.

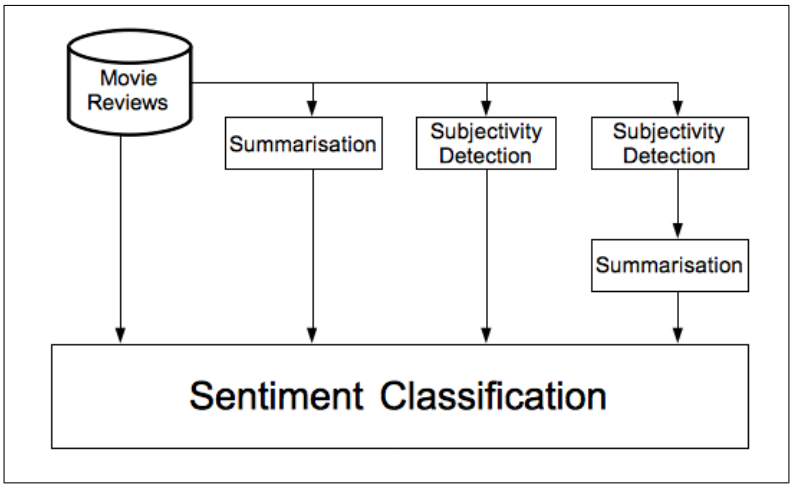

Fig. 1. Pipeline of the review summarisation and classification

3.1 Sentiment Classification

Sentiment classification is a text classification task, where a label indicates the polarity of the document rather than its topic. The task can be approached from different points of view. For example, identifying the overall sentiment of a document is different from mining the polarity of individual aspects like soundtrack, plot, etc. In this paper, only the polarity of the document as a whole is considered, i.e. whether the overall recommendation of a review is positive or negative.

Traditional machine learning approaches can be applied for this classification task. Specifically, Naive Bayes (NB) and Support Vector Machine (SVM) classifiers are considered, using unigram-presence as features. The feature selection for NB is based on document frequency, being a commonly used selection strategy.

3.2 Extractive Summarisation

In order to produce different kinds of extractive summaries, different sentence selection techniques are applied. Notice that using unigrams as features for the classification, rebuilding the original order of the sentences is necessary only when a further summarisation step, which considers sentence position or proximity, is performed.

The considered techniques are the following:

- Luhn's traditional approach, as representative of statistical approaches;
- positional approaches, based on the intuition that the location of the sentence within the document reflects its significance;
- subjectivity detection, used to filter out sentences which do not express opinions;
- combinations of subjectivity detection with the other approaches.

Luhn's approach Firstly, the traditional Luhn's approach [6] is used to score the sentences according to their significance. The top N sentences are selected to create the summary. The results for this approach are labelled as *Luhn-N*, where N is the number of sentence used to create the summary. The significance score of a sentence is based on clustering of sentence tokens using a distance threshold (5 is the distance used in this paper). For each cluster, the score is computed taking the ratio between the square of the number of significant words in the cluster, over the total number of words in the cluster. The significant words are chosen according to their frequency, i.e. the terms with higher tf, excluding stop words, are considered significant. The significance score for a sentence will be the maximum score for any of its clusters.

Position-based approaches A second family of summarisers is built on top of an empirical observation: often reviewers tend to summarise their overall feeling in a sentence or in a short paragraph, placed either at the beginning or at the end of the review. In this case, a summary can be created simply selecting the N opening sentences, or the N closing sentences. Results for these approaches are labelled as *First-N* and *Last-N*, respectively.

Subjectivity detection The previous approaches do not take into account the subjective nature of the documents under analysis. To overcome this issue, the aforementioned classification techniques can be used to identify and filter subjective sentences. A specific data-set, described in Section 4, is used to train the classifiers. Filtering out the objective sentences and aggregating only the subjective ones can already be seen as a summarisation approach. The average compression rate of the data under analysis is around 60%. Results for this approach are labelled as *Subjective-Full*.

Summarising subjective extracts In order to further increase the compression rate, and to take into account subjectivity, one of the first two approaches can be applied to the subjective extracts. In the results, this family of approaches is labelled as follows: *Subjective-Luhn-N* for the summaries produced using Luhn's approach on the subjective sentences, *Subjective-First-N* and *Subjective-Last-N* for the summaries based on the subjective sentence positions. Again, N represents the number of selected sentences.

4 Experimental Study

The evaluation of summarisation systems is a research issue in itself, and different intrinsic evaluation approaches have been proposed over the years [7]. Since the purpose of this work is observing how the use of summarisation techniques can help the sentiment classification task, we do not evaluate the summaries with traditional methods like ROUGE [5] or Pyramid [8], nor we look for linguistic quality. The evaluation is performed with respect to the polarity classification, i.e. a good summary is ideally able to carry the same polarity of the full document. Full text reviews and summaries are classified according to their overall polarity.

4.1 Experimental Setup

For the subjectivity detection, a data-set of subjective and objective sentences is used to train the classifiers [9]. This data-set contains 5000 subjective sentences, taken from RottenTomatoes snippets, and 5000 objective sentences, which are taken from IMDb plots. The main idea behind the creation of the subjectivity data-set consists in assuming that the review snippets from RottenTomatoes contain only opinionated sentences, while the movie plots taken from IMDb contain non-opinionated, and hence objective, sentences. Firstly, the classifiers are tested on the subjectivity data-set, using a five-folding cross-validation approach. The micro-averaged F_1 results are not significantly different (88.85 for NB vs. 88.68 for SVM). The classifiers can be considered reliable enough for the subjectivity detection task which leads to the generation of subjective extracts.

The sentiment classification has been evaluated on the movie review data-set firstly used in [9], containing reviews taken from IMDb[2] and annotated as positive or negative. The data-set contains 2000 documents, evenly distributed between the two classes.

4.2 Results and Discussion

Table 1 reports the results of the micro-averaged F_1 scores on the review data-set. This evaluation measure is chosen as it is one of the most commonly used in text classification [12]. The macro-averaged results are not reported as they are very similar to the micro-averaged ones, given the data-set is well balanced, i.e. the two classes contain the same number of document.

Table 1. The micro-averaged F_1 results of sentiment classification

	NB	SVM		NB	SVM
Full Review	83.31	87.10	Subjective-Full	84.61	86.82
Luhn-1	70.12	70.28	Subjective-Luhn-1	71.02	70.50
Luhn-3	75.47	74.96	Subjective-Luhn-3	74.92	74.91
First-1	68.94	68.82	Subjective-First-1	69.33	68.90
Last-1	70.61	70.49	Subjective-Last-1	70.90	71.15
First-3	70.81	70.43	Subjective-First-3	71.12	71.07
Last-3	75.58	76.57	Subjective-Last-3	75.49	76.26

The first observation is that statistics and positional summarisation approaches do not provide any improvement to the sentiment classification results. On the contrary, the performances are substantially worse for both NB and SVM. The explanation behind this behaviour is that these approaches are not explicitly opinion-oriented, so they are not able to capture the sentiment behind a review.

[2] http://www.imdb.com

The quality of sentiment classification for subjective extracts is instead in line with the full review classification. More precisely, the classification of subjective extracts through NB achieves a 1.5% better result compared to the classification of full text. On the SVM side, the classification of subjective extracts is performed slightly worse than the classification of full text. In other words, the subjectivity detection step preserves the most important information about polarity, and this aspect is captured by both classifiers. In order to double check this finding, experiments on objective extracts classification have been also performed. The objective sentences have been aggregated, building the counterparts of the subjective extracts. The micro-averaged F_1 values for the objective extracts classification were below 75% for both classifiers, hence significantly worse than both the full review and subjective extract classification. When further summarisation is performed on the subjective extracts, the results drop again. On the two sides of Table 1, we can observe a similar behaviour between summaries created from the full text and summaries created from the subjective extracts.

As further analysis, we also examine the classification of the summaries with respect to the full documents. In other words, we check if a full text and its respective summary are classified under the same label, without considering whether this is the correct answer or not. In 91% of the cases, the subjective summaries are assigned to the same label of the correspondent full text. For all the other summarisation approaches, this value drops below 80%, and in some cases below 70%. This is a further evidence of the connection between subjectivity and polarity.

5 Conclusion and Future Work

This paper has investigated the use of extractive summarisation in the context of sentiment classification. Experiments using NB and SVM classifiers have been carried out on a movie review data-set, in order to classify documents according to their polarity. Different summarisation techniques have been applied to the reviews, with the purpose of building summaries which capture the polarity of the respective original documents. Sentence extraction techniques based on statistical or positional approaches fail to capture the subjectivity of the review, and hence are inadequate to represent the sentiment of the document. On the contrary, using subjectivity detection to build subjective extracts produces results which are comparable to the full text classification. Further summarisation on top of subjectivity detection, again fail to capture the polarity of documents, as more opinion-oriented approaches needed. Showing a subjective extract instead of the full text, a potential user would only need to read 60% of a review, or even less, in order to understand its polarity.

For the future, we intend to investigate the use of knowledge extraction techniques, in order to identify entities and relationships between entities. The benefits of this approach include the opportunity of analysing opinions at a finer granularity, i.e. not only classifying the overall polarity, but also the polarity with respect to individual aspects of movies or products. This can be extended to multi-document summarisation, and would lead to the generation of personalised summaries.

References

1. P. Beineke, T. Hastie, C. Manning, and S. Vaithyanathan. Exploring sentiment summarization. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (AAAI tech report SS-04-07)*, 2004.
2. S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G.A. Reis, and J. Reynar. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era*, 2008.
3. J.M. Chenlo and D.E. Losada. Effective and efficient polarity estimation in blogs based on sentence-level evidence. In *Proceedings of the 20th ACM international conference on information and knowledge management*, pages 365–374. ACM, 2011.
4. M. Hu and B. Liu. Mining opinion features in customer reviews. In *Proceedings of the National Conference on Artificial Intelligence*, pages 755–760. AAAI, 2004.
5. C.Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, pages 25–26, 2004.
6. H.P. Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.
7. A. Nenkova and K. McKeown. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233, 2011.
8. A. Nenkova and R.J. Passonneau. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL*, pages 145–152, 2004.
9. B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 271–278. Association for Computational Linguistics, 2004.
10. B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
11. M. Potthast and S. Becker. Opinion Summarization of Web Comments. In C. Gurrin et al., editor, *Proceedings of the 32nd European Conference on Information Retrieval, ECIR 2010*, volume 5993 of *Lecture Notes in Computer Science*, pages 668–669, 2010.
12. Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.
13. D. Shen, Z. Chen, Q. Yang, H.J. Zeng, B. Zhang, Y. Lu, and W.Y. Ma. Web-page classification through summarization. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 242–249. ACM, 2004.
14. P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.
15. L. Zhuang, F. Jing, and X.Y. Zhu. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50. ACM, 2006.