Summarizing Opinions in Blog Threads [⋆]

Alexandra Balahur[a,b], Mijail Kabadjov[b], Josef Steinberger[b], Ralf Steinberger[b], and Andrés Montoyo[a]

[a] Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante
Apartado de Correos 99, E-03080, Alicante , Spain
{abalahur, montoyo}@dlsi.ua.es
[b] Joint Research Centre, European Commission,
Via E. Fermi 2749, 21027, Ispra (VA), Italy
{mijail.kabadjov, josef.steinberger, ralf.steinberger}@jrc.ec.europa.eu

Abstract. In this paper we present an approach to summarizing positive and negative opinions in blog threads. We first run a sentiment analysis system and consequently pass its output through a standard LSA-based text summarization system. Further on, we evaluate our approach and present the results obtained, which we believe are promising in the context of multi-document text summarization. Finally, we discuss the main issues in applying standard text summarization techniques to the slightly different task of summarizing opinions in blog threads.

Keywords: Sentiment Analysis, Opinion Mining, Text Summarization

1 Introduction

Recent years have brought about an important shift in the way objective and subjective information are regarded and impact society and its individuals. It is no longer only the factual content (of news), but rather what people feel about it, that influences decisions taken daily. Supported by the fast development of the Internet and the Web 2.0 technologies, with the predominant presence of social networks, forums, "blogging" and reviewing as world-wide phenomena, exchanging views and debating on real-life related issue has reached a global scale. People express and search for opinions on blogs, forums, in reviews and comments – leading to the creation of extensive quantities of data that cannot be manually processed, although their analysis (discovery of opinions, their classification into positive and negative), could be useful to a high diversity of entities (potential customers, companies, public figures and institutions etc.), for a large variety of tasks (opinion analysis for marketing, sociological or political studies, decision support etc.). Automatic systems are thus needed to help resolve the issue of large-scale data analysis. Example of such a system would be one that is analyzing subjective data expressed on the Web on a certain topic and presenting the potential users with short summaries of the main points of view expressed, depending on whether or not they are in favor or against the topic.

This paper presents an approach to building such a system that is able to output summaries of the points of view expressed in blog threads. The rest of the paper is organized as follows: in section 2 we briefly discuss related work; in section 3 we present our approach to opinion summarization and give details of the corpus we used to develop and evaluate our system; next, we present our experimental results and discuss the main issues (sec. 4); and finally, we conclude the paper and give pointers to future work.

[⋆] Special thanks to the EMM team for supporting this research. We also thank Ester Boldrini for her annotations of the corpus used in the experiments presented.

2 Related Work

Whilst there is abundant literature on text summarization (Kabadjov *et al.*, 2009; Hovy, 2005; Erkan and Radev, 2004; Gong and Liu, 2002) and sentiment analysis (Balahur *et al.*, 2009a; Pang and Lee, 2008; Riloff *et al.*, 2005), there is still limited work at the intersection of these two areas (Stoyanov and Cardie, 2006).

Initial research in opinion mining concentrated on news texts. Wiebe (1994) defines subjectivity based on Quirks idea of "private states" (states that are not open to verification) and distinguishes between objectivity and subjectivity on this criteria. Consequently, based on this definition, the Multi-Perspective Question Answering (MPQA) annotation schema and corpus were created over news texts, distinguishing between the subjective/objective speech, as well as the polarity of text spans (Wiebe *et al.*, 2005). Subsequently, different authors show that this initial discrimination is crucial for the sentiment task, improving results obtained when using only polarity classification for sentence-level opinion mining (Pang and Lee, 2004), as part of Opinion Information Retrieval (last three editions of the TREC Blog tracks, the TAC 2008 competition), Information Extraction (Riloff *et al.*, 2005) and Question Answering (QA) (Stoyanov *et al.*, 2004) systems. Once this discrimination is done, or in the case of texts containing only or mostly subjective language (such as e-reviews), opinion mining becomes a polarity classification task.

3 Opinion Summarization

In our opinion summarization experiments we adopt a standard approach by employing in tandem a sentiment classification system and a text summarizer. The output of the former is used to divide the sentences in the blog threads into three groups: sentences containing positive sentiment, sentences containing negative sentiment and neutral or objective sentences. Then the positive and the negative sentences are passed on to the summarizer separately to produce one summary for the positive posts and another one for the negative ones. Thus, for each blog thread we produce two summaries: one positive and one negative.

Clearly, integrating two systems in tandem results in a composite system in which errors from the first system are propagated onto the next, and consequently reducing the overall system performance in comparison to the case where the two constituting systems are evaluated in isolation. However, we believe that by using this type of architecture one obtains a fairer and more realistic perspective on the performance of such system in real applications.

We discuss in more detail the sentiment classification and the summarization systems below, but first we describe the corpus we used to develop and evaluate our opinion summarizer.

3.1 Sentiment Annotated Corpus

3.1.1 Annotation Process The corpus we employed in this study is a collection of 51 blog threads extracted from the Web. The motivation for using such a small corpus is that it can easily be annotated manually, and at the same time it is sufficient for a preliminary study of the phenomena found in such types of texts. The blog threads are written in English and have the same structure: an initial post by the blog author, containing a piece of news and the author's opinion on it, followed by a set of comments by other bloggers and/or the author in reply to the initial blog post or one of the previous comments. The topics they cover are: economy, science and technology, cooking, society and sport. In most cases, the comment posts are the most subjective texts even if the initial intervention the author can also include expressions of opinions. Blogs can also contain multi-modal information, but we decided to take into account only the text. In our blog corpus annotation, we indicated the url from which the thread was extracted, we then included the initial annotated piece of news and the labeled user comments. Table 1 shows the average and the total number of posts, of words in the news, of the number of words in posts and, finally, of words both in news and in posts included in the corpus at hand.

Table 1: Corpus size

	N. Posts	N. Words for news	N. Words for post	Total words
Total	1829	72.995	226.573	299.568
Average	33.87	1351.75	4195.79	5547.55

As it can be seen in Table 1, the corpus at hand is rather small. However, its size allows for an in-depth analysis of the phenomena such texts contain and the study of the success and failure factors in our approach. In order to create a Gold Standard on the basis of which we can evaluate our sentiment analysis and summarization systems, we labeled the corpus of blog threads using some of the EmotiBlog (Balahur *et al.*, 2009b) elements presented in Table 2.

Table 2: Annotated elements

Element	Attribute
Polarity	Positive, negative
Level	Low, medium, high
Source	name
Target	name

As we can see in Table 2, only the annotation elements that were relevant to our analysis were chosen. Each of the blog threads were classified according to the topics they covered. Thus, for each of the blogs considered, we annotated the main and secondary topics it covers. Within the comment post, we firstly discriminated between objective (sentences that were either factual, neutral or subjective, but not related to the topics assigned) and subjective sentences (positive or negative opinion sentences related to the topics assigned). Subsequently, we took into consideration only the subjective sentences with the elements presented in the table. Each of the elements indicated in the table above has been selected because they provide important information that is relevant to the task at hand. The polarity has the function of indicating if the opinion expressed in the sentence is positive of negative. The 'level' element shows the intensity of the opinion expressed; its possible values are low, medium or high. Finally, the source of the discourse is specified in order to be able to detect the source and the target of the sentence, as well as which of the topics included in the thread the sentence refers to. The result of the annotation process is a gold standard which will be used to evaluate the sentiment analysis system and the final generated summaries.

Figure 1 is an example of annotation. As it can be noticed, more than one topic per blog thread is indicated, since the presence of multiple subjects for discussion is something typical to the blogosphere. Since only sentences containing opinions relevant to the topics considered are annotated, the topic delimitation is an important component of the labeled data. In the example presented, the main topic is the economic situation, while the secondary ones are the government and banks. After having defined the topics, the first paragraph contains objective information and thus, we do not label it; we therefore annotate the following sentence that contains subjective information. As you can see, the economic crisis is the target. Finally, the polarity of the sentence is negative, the intensity level of this polarity is medium and the author is Cynicus Economicus.

3.1.2 Annotation problems During the annotation process we faced some difficulties, to which we tried proposing possible solutions. The first obstacle we detected consisted in finding the topic of each blog. A phenomenon that we detected as particular to blogs is that very often there is a mixture of topics present in the argumentation line. Moreover, people add copy and pastes from newspaper articles or other types of media in order to support their ideas. Furthermore, it is very usual that the author of the new writes about a topic, but during the discussion in the blog, people change the topic of conversation. In order to overcome these problems, we decided to insert more than one topic, given that they are relevant to the global discourse. The judgement of the sentence relevance was done based on the subjectivity it contains and the relevance it has for the topic in question.

```
<topic>economic situation</topic>
<topic2>government</topic2>
<topic3>banks</topic3>
<new>Saturday, May 9, 2009. My aim in this blog has largely been to
 give my best and most rational perspective on the reality of the
 economic situation. I have tried (and I hope) mostly succeeded in
 avoiding emotive and partisan viewpoints, and have tried as far as
 possible to see the actions of politicians as misguided. Of late,
 that perspective has been slipping, for the UK, the US and also
 for Europe.
 <phenomenon gate:gateId="1" target="economic crisis" degree1="medium"
  category="phrase" source="Cynicus Economicus" polarity1="negative">I
  think that the key turning point was the Darling budget, in which
  the forecasts were so optimistic as to be beyond any rational
  belief</phenomenon>
</new>
```

Figure 1: Example of labeling

3.2 Sentiment Analysis

The first step we took in our approach was to determine the opinionated sentences, assign each of them a polarity (among 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). Given that we are faced with the task of classifying opinion in a general context, we employed a simple, yet efficient approach, presented in (Balahur *et al.*, 2009c). At the present moment, there are different lexicons for affect detection and opinion mining. In order to have a more extensive database of affect-related terms, in the following experiments we used WordNet Affect (Strapparava and Valitutti, 2004), SentiWordNet (Esuli and Sebastiani, 2006), MicroWNOp (Cerini *et al.*, 2007). Each of the employed resources were mapped to four categories, which were given different scores: positive (1), negative (-1), high positive (4) and high negative (-4). As shown in (Balahur *et al.*, 2009c), these values performed better than the usual assignment of only positive (1) and negative (-1) values. First, the score of each of the blog posts was computed as sum of the values of the words identified; a positive score leads to the classification of the post as positive, whereas a final negative score leads to the system classifying the post as negative. Subsequently, we performed sentence splitting using Lingpipe and classified the obtained sentences according to their polarity, by adding the individual scores of the affective words identified. As it has been shown in (Balahur *et al.*, 2009c), some resources tend to over classify positive or negative examples. Thus, we have used the combined resources, which have proven to classify in a more balanced manner (Balahur *et al.*, 2009c). The measure of the intensity of the scores can also be used as an indication of the sentence importance and can thus constitute a criterion for summarization, as shown in (Balahur *et al.*, 2008).

3.3 LSA-based Text Summarization

Originally proposed by (Gong and Liu, 2002) and later improved by (Steinberger and Ježek, 2004), this approach first builds a term-by-sentence matrix from the source, then applies a powerful statistical technique for matrix decomposition called Singular Value Decomposition (SVD) and finally uses the resulting matrices to identify and extract the most salient sentences. The idea behind this is that the SVD finds the latent (orthogonal) dimensions, which in simple terms correspond to the different topics discussed in the source (fine-tuning of the model is necessary, though, for optimal performance).

More formally, we first build matrix $A = [A_1 \dots A_n]$, where each column $A_i = [a_{1i} \dots a_{ni}]^T$ represents the weighted term-frequency vector of sentence i in a given document. Each element in this vector is defined as $a_{ji} = L(t_{ji}) \cdot G(t_{ji})$, where t_{ji} denotes the frequency with which term j occurs in sentence i, $L(t_{ji})$ is the local weight for term j in sentence i, and $G(t_{ji})$ is the global weight for term j in the whole document. We use a binary local weight and an entropy-based global weight (for more details see (Steinberger *et al.*, 2007)).

Once the matrix A is built, it is decomposed via Singular Value Decomposition (SVD) defined as $A = U\Sigma V^T$. Due to space constraints, details of how sentences are extracted using matrices V^T and Σ are omitted here (see (Steinberger *et al.*, 2007) for details on that).

4 Experimental Results

We first discuss the performance of the sentiment recognition system followed by the summarization performance.

Performance results of the sentiment analysis are shown in Table 3.

Table 3: Sentiment analysis performance

System	Precision	Recall	F1
$Sent_{neg}$	0.98	0.54	0.69
$Sent_{pos}$	0.07	0.69	0.12

We have also analyzed in depth the results of 14 blog threads in order to assess the quality of the opinion mining, independently of the topic relevance of the sentences classified. As only sentences that were relevant to the topic in question were labeled in the gold standard, we also assessed the sentiment of the sentences that were not annotated.

Table 4: Sentiment analysis performance in detail

Thread No.	Pos sent				Neg sent				Asserted Pos	Asserted Neg
	Total	O-P	O-N	O-O	Total	O-P	O-N	O-O		
1	5	0	1	2	4	0	2	0	2	2
2	5	0	0	1	5	1	2	1	1	1
3	4	1	2	1	4	0	2	0	0	2
4	4	2	1	1	4	0	2	1	0	1
6	5	1	4	0	5	0	4	1	0	1
8	5	4	0	1	6	0	3	2	0	1
11	4	3	0	1	6	3	3	0	0	0
12	4	0	3	1	6	0	3	1	0	2
14	5	1	3	1	5	0	4	0	0	1
15	5	2	2	1	5	0	3	2	0	0
16	4	2	2	0	5	1	2	0	0	2
18	5	0	3	2	4	1	0	3	0	0
30	7	3	3	1	4	0	2	0	0	2
31	5	2	3	0	5	1	3	0	0	1

As we can observe from the results presented in Table 3, the system employed had a relatively high recall and a low precision, meaning that the sentences that were classified as positive or negative by the system were either wrongfully classified or they were annotated in the Gold Standard as being objective. However, the high recall suggests that the system, although simple, is capable of distinguishing subjective sentences from objective ones. As we can observe from the results presented in Table 4, the opinion mining system performed well as far as polarity classification was concerned. Thus, although relatively few of the sentences in the summary are also present in the Gold Standard as far as polarity and sentence importance are concerned, the sentences were classified well, especially for the negative class (correlation between Negative and O-N and Positive and O-P). Thus, improvement can be achieved over these results by adding a topic detection component. It can also be noticed that the system has the tendency to overclassify sentences as being negative, a fault which we attribute to the fact that in our approach we do not contemplate negations and the fact that the resources used contain, in their original form, word senses, and we do not perform any word sense disambiguation. Also, most of the resources used for sentiment detection have a large number of negative terms and a significantly lower number of positive ones.

Performance results of the summarizer are shown in Table 5 below. We used the standard ROUGE evaluation (Lin and Hovy, 2003) which has been also used for the *Text Analysis Conferences*. We include the usual ROUGE metrics: R_1 is the maximum number of co-occurring unigrams, R_2 is the maximum number of co-occurring bigrams, R_{SU4}, is the skip bigram measure with the addition of unigrams as counting unit, and finally, R_L is the longest common subsequence measure (Lin, 2004). In all cases we present the average F1 score for the given metric and within parenthesis the 95% confidence intervals.

Table 5: Summarization performance

System	R_1	R_2	R_{SU4}	R_L
$Sent + Summ_{neg}$	0.22 (0.18 - 0.26)	0.09 (0.06 - 0.11)	0.09 (0.06 - 0.11)	0.21 (0.17 - 0.24)
$Sent + Summ_{pos}$	0.21 (0.17 - 0.26)	0.05 (0.02 - 0.09)	0.05 (0.02 - 0.09)	0.19 (0.16 - 0.23)
$Summ_{TAC08}$	0.348	0.081	0.12	–

There are three rows in Table 5: the first one ($Sent + Summ_{neg}$) is the performance of the LSA summarizer on the negative posts, the second one ($Sent + Summ_{pos}$) presents the performance of the LSA summarizer on the positive posts and the third one is only included here for reference and corresponds to the official performance of an LSA summarizer using the same method as ours at the 2008 *Text Analysis Conference* Summarization track (TAC08). The latter provide a reasonable context for the results on opinion summarization (i.e., the top two rows). [1]

The first thing to note from Table 5 is that the performance on negative posts is better, though, being within the 95% confidence intervals, the difference cannot be considered statistically significant. One possible reason for the slightly better performance on the negative posts is that the sentiment recognition system is more accurate with negative sentiment than with positive.

The other observation we make is that the TAC08 summarization performance is either close or within the 95% confidence intervals. It is worth noting that the LSA summariser employing the same method as our LSA summarizer ranked in the top 20% summarization systems at the TAC 2008 competition. Additionally, the same LSA method has already been improved upon by incorporating higher level semantic information such as coreference (Steinberger *et al.*, 2007), and hence, applying the same method in our context would also potentially translate in performance improvement. In the light of this, we believe the performance results attained are promising.

[1] We note, however, that the results on our corpus are not directly comparable with those of TAC08, since the data sets are different.

The main problem we encountered was that the LSA-based summarization method we adopted was originally designed to work with grammatical sentences from news articles. And in our case blog posts are often composed of ungrammatical sentences and also there is a high number of strange characters such as `:-), ;), :-(` etc. which make the blog data much nosier and harder to process than the standard data sets traditionally used for summarization evaluation. However, in our case the LSA method, being a statistical method, proved to be quite robust to variations in the input data and most importantly to the change of domain.

5 Conclusion and Future Work

In this paper we demonstrated that an obvious approach to the task of opinion summarization by combining two separate systems, one for sentiment classification and one for text summarization, is, indeed, a feasible approach to that task and yields reasonable performance results on a specialized blog corpus annotated for sentiment and summarization.

In future work we intend to exploit higher level semantic information such as entities and taxonomies as in (Steinberger *et al.*, 2007; Kabadjov *et al.*, 2009).

References

Balahur, A., E. Boldrini, A. Montoyo, and P. Martínez-Barco. 2009a. Cross-topic opinion mining for real-time human-computer interaction. In *Proceedings of ICEIS 2009 Conference*.

Balahur, A., E. Boldrini, A. Montoyo, and P. Martínez-Barco. 2009b. Emotiblog: an annotation scheme for emotion detection and analysis in non-traditional textual genres. In *Proceedings of the Data Mining Conference (DMIN)*.

Balahur, A., E. Lloret, O. Ferrández, A. Montoyo, M. Palomar, and R. Muñoz. 2008. The DLSIUAES team's participation in the TAC 2008 tracks. In of N. I. Standards and Technology. , editors, *Proceedings of the Text Analysis Conference*, Gaithersburg, MD.

Balahur, A., R. Steinberger, E. van der Goot, and B. Pouliquen. 2009c. Opinion mining from newspaper quotations. In *Proceedings of the Workshop on Intelligent Analysis and Processing of Web News Content at the IEEE / WIC / ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT)*.

Cerini, S., V. Compagnoni, A. Demontis, M. Formentelli, and G. Gandini. 2007. Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In A. Sansò, editor, *Language resources and linguistic theory: Typology, second language acquisition, English linguistics*. Franco Angeli, Milano, IT.

Erkan, G. and D. R. Radev. 2004. LexRank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.

Esuli, A. and F. Sebastiani. 2006. SentiWordNet: A publicly available resource for opinion mining. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Italy.

Gong, Y. and X. Liu. 2002. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of ACM SIGIR*, New Orleans, US.

Hovy, E. H. 2005. Automated text summarization. In R. Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, pages 583–598. Oxford University Press, Oxford, UK.

Kabadjov, M. A., J. Steinberger, B. Pouliquen, R. Steinberger, and M. Poesio. 2009. Multilingual statistical news summarisation: Preliminary experiments with english. In *Proceedings of the Workshop on Intelligent Analysis and Processing of Web News Content at the IEEE / WIC / ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT)*.

Lin, C.-Y. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain.

Lin, C.-Y. and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL*, Edmonton, Canada.

Pang, B. and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 271–278, Barcelona, Spain. Association for Computational Linguistics.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Riloff, E., J. Wiebe, and W. Phillips. 2005. Exploiting subjectivity classification to improve information extraction. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI)*.

Steinberger, J. and K. Ježek. 2004. Text summarization and singular value decomposition. In *Proceedings of the 3rd ADVIS conference*, Izmir, Turkey.

Steinberger, J., M. Poesio, M. A. Kabadjov, and K. Ježek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43(6):1663–1680. Special Issue on Text Summarisation (Donna Harman, ed.).

Stoyanov, V. and C. Cardie. 2006. Toward opinion summarization: Linking the sources. In *Proceedings of the COLING-ACL Workshop on Sentiment and Subjectivity in Text*, Sydney, Australia. Association for Computational Linguistics.

Stoyanov, V., C. Cardie, D. Litman, and J. Wiebe. 2004. Evaluating an opinion annotation scheme using a new multi-perspective question and answer corpus. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Strapparava, C. and A. Valitutti. 2004. WordNet-Affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086, Lisbon, Portugal.

Wiebe, J. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

Wiebe, J., T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.