New Features for Sentiment Analysis: Do Sentences Matter?

Gizem Gezici[1], Berrin Yanikoglu[1], Dilek Tapucu[1,2], and Yücel Saygın[1]

[1] Faculty of Engineering and Natural Sciences, Sabancı University, Istanbul, Turkey
{gizemgezici,berrin,dilektapucu,ysaygin}@sabanciuniv.edu
[2] Dept. of Computer Engineering, Izmir Institute of Technology, Izmir, Turkey

Abstract. In this work, we propose and evaluate new features to be used in a word polarity based approach to sentiment classification. In particular, we analyze sentences as the first step before estimating the overall review polarity. We consider different aspects of sentences, such as length, purity, irrealis content, subjectivity, and position within the opinionated text. This analysis is then used to find sentences that may convey better information about the overall review polarity. The TripAdvisor dataset is used to evaluate the effect of sentence level features on polarity classification. Our initial results indicate a small improvement in classification accuracy when using the newly proposed features. However, the benefit of these features is not limited to improving sentiment classification accuracy since sentence level features can be used for other important tasks such as review summarization.

Keywords: sentiment analysis; sentiment classification; polarity detection; machine learning

1 Introduction

Sentiment analysis aims to extract the opinions indicated in textual data enabling us to understand what people think about specific issues by analyzing large collections of textual data sources such as personal blogs, review sites, and social media. An important part of sentiment analysis boils down to a classification problem, i.e., given an opinionated text, classifying it as positive or negative polarity and Machine Learning techniques have already been adopted to solve this problem.

Two main approaches for sentiment analysis are lexicon-based and supervised methods. The lexicon-based approach calculates the semantic orientation of words in a review by obtaining word polarities from a lexicon such as the SentiWordNet [5]. While the SentiWordNet [5] is a domain-independent lexicon, one can use a domain-specific lexicon whenever available since domain-specific lexicons better indicate the word polarities in that domain (e.g. the word "small" has a positive connotation in cell phone domain; while it is negative in hotel domain).

Supervised learning approaches use machine learning techniques to establish a model from a large corpus of reviews. The set of sample reviews form the training data from which the model is built. For instance in [16] [21] , researchers use the Naive Bayes algorithm to separate positive reviews from negative ones by learning the probability distributions of the considered features in the two classes. While supervised approaches are typically more successful, collecting a large training data is often a problem.

Word-level polarities provide a simple yet effective method for estimating a review's polarity, however, the gap from word-level polarities to review-level polarity is too big. To bridge this gap, we propose to analyze word-polarities within sentences, as an intermediate step.

The idea of sentence level analysis is not new. Some researchers approached the problem by first finding subjective sentences in a review, with the hope of eliminating irrelevant sentences that would generate noise in terms of polarity estimation [13], [24]. Yet another approach is to exploit the structure in sentences, rather than seeing a review as a bag of words [8][11][15]. For instance in [8], conjunctions were analyzed to obtain the polarities of the words that are connected with the conjunct. In [9],[14] researchers focused on sentence polarities separately, again to obtain sentence polarities more correctly, with the goal of improving review polarity in turn. The first line polarity has also been used as a feature by [24].

Similar to [24], this work is motivated by our observation that the first and last lines of a review are often very indicative of the review polarity. Starting from this simple observation, we formulated more sophisticated features for sentence level sentiment analysis. In order to do that, we performed an in-depth analysis of different sentence types. For instance, in addition to subjective sentences, we defined pure, short, and no irrealis sentences.

We performed a preliminary evaluation using the TripAdvisor dataset to see the effect of sentence level features on polarity classification. Throughout the evaluation, we observed a small improvement in classification accuracy due to the newly proposed features. Our initial results showed that the sentences do matter and they need to be explored in larger and more diverse datasets such as blogs. Moreover, the benefit of these features is not limited to improving sentiment classification accuracy. In fact, sentence level features can be used to identify the essential sentences in the review which could further be used in review summarization.

Our paper is organized as follows: Section 2 presents our taxonomy of sentiment analysis features, together with the newly proposed features. Section 3 describes the sentence level analysis for defining the features. Section 4 describes the tools and methodology for sentiment classification together with the experimental results and error analysis. Finally, in Section 5 we draw some conclusions and propose future extension of this work.

2 Taxonomy and Formulation of the New Features

We define an extensive set of 19 features that can be grouped in four categories: (1) basic features, (2) features based on subjective sentence occurrence statistics, (3) delta-tf-idf weighting of word polarities, and (4) sentence-level features. These features are listed in Table 1 and using the notations given below and some basic definitions provided in Table 2, they are defined formally in Tables 3-7.

Table 1. Summary Feature Descriptions for a Review R

Group Name	Feature	Name
Basic	F_1	Average review polarity
	F_2	Review purity
Occurrence of Subjective Words	F_3	Freq. of subjective words
	F_4	Avg. polarity of subj. words
	F_5	Std. of polarities of subj. words
$\Delta TF * IDF$	F_6	Weighted avg. polarity of subj. words
	F_7	Scores of subj. words
Punctuation	F_8	# of Exclamation marks
	F_9	# of Question marks
Sentence Level	F_{10}	Avg. First Line Polarity
	F_{11}	Avg. Last Line Polarity
	F_{12}	First Line Purity
	F_{13}	Last Line Purity
	F_{14}	Avg. pol. of subj. sentences
	F_{15}	Avg. pol. of pure sentences
	F_{16}	Avg. pol. of non-irrealis sentences
	F_{17}	$\Delta TF * IDF$ weighted polarity of first line
	F_{18}	$\Delta TF * IDF$ scores of subj. words in the first line
	F_{19}	Number of sentences in review

A review R is a sequence of sentences $R = S_1S_2S_3...S_M$ where M is the number of sentences in R. Each sentence S_i in turn is a sequence of words, such that $S_i = w_{i1}w_{i2}...w_{iN(i)}$ where $N(i)$ is the number of words in S_i. The review R can also be viewed as a sequence of words $w_1..w_T$, where T is the total number of words in the review.

In Table 2, subjective words (SBJ) are defined as all the words in SentiWordNet that has a dominant negative or positive polarity. A word has dominant positive and negative polarity if the sum of its positive and negative polarity values is greater than 0.5 [23]. $SubjW(R)$ is defined as the most frequent subjective words in SBJ (at most 20 of them) that appear in review R. For a sentence $S_i \in R$, the average sentence polarity is used to determine subjectivity of that sentence. If it is above a threshold, we consider the sentence as subjective, forming $subjS(R)$. Similarly, a sentence S_i is pure if its purity is greater than a fixed threshold τ. We experimented with different values of τ and for evaluation we used $\tau = 0.8$. These two sets form the $subS(R)$ and $pure(R)$ sets respectively.

We also looked at the effect of first and last sentences in the review, as well as sentences containing irrealis words. In order to determine irrealis sentences, the existence of the modal verbs 'would', 'could', or 'should' is checked. If one of these modal verbs appear in the sentence then these sentences are labeled as irrealis similar to [17].

Table 2. Basic definitions for a review R

M	the total number of sentences in R
T	the total number of words in R
SBJ	set of known subjective words
$subjW(R)$	set of most frequent subjective words from SBJ, in R (max 20)
$subjS(R)$	set of subjective sentences in R
$pure(R)$	set of pure sentences in R
$nonIr(R)$	set of non-irrealis sentences in R

2.1 Basic Features

For our baseline system, we use the average word polarity and purity defined in Table 3. As mentioned before, these features are commonly used in word polarity based sentiment analysis. In our formulation $pol(w_j)$ denotes the dominant polarity of w_j of R, as obtained from SentiWordNet, and $|pol(w_j)|$ denotes the absolute polarity of w_j.

Table 3. Basic Features for a review R

F_1	Average review polarity	$\frac{1}{T}\sum_{j=1..T} pol(w_j)$
F_2	Review purity	$\frac{\sum_{j=1..T} pol(w_j)}{\sum_{j=1..T} \|pol(w_j)\|}$

2.2 Frequent Subjective Words

The features in this group are derived through the analysis of subjective words that frequently occur in the review. For instance, the average polarity of the most frequent subjective words (feature F_4) aims to capture the frequent sentiment in the review, without the noise coming from *all* subjective words.

The features were defined before in some previous work [4]; however, to the best of our knowledge, they considered all words, not specifically subjective words.

2.3 Δtf*idf Features

We compute the $\Delta tf*idf$ scores of the words in SentiWordNet [5] from a training corpus in the given domain, in order to capture domain specificity [12]. For a word w_i, $\Delta tf*idf(w_i)$ is defined as $\Delta tf*idf(w_i) = tf*idf(w_i,+) - tf*idf(w_i,-)$.

Table 4. Features Related to Frequency and Subjectivity

F_3	Freq. of subjective words	$\|subjW(R)\|/\|R\|$
F_4	Avg. polarity of subj. words	$\frac{1}{\|subjW(R)\|}\sum_{w_j \in subjW(R)} pol(w_j)$
F_5	Stdev. of polarities of subj. words	$\sqrt{\frac{1}{\|subjW(R)\|}\sum_{w_j \in subjW(R)}(pol(w_j) - F_4)^2}$

If it is positive, it indicates that a word is more associated with the positive class and vice versa, if negative. We computed these scores on the training set which is balanced in the number of positive and negative reviews.

Then, we sum up the $\Delta tf * idf$ scores of these words (feature F_6). By doing this, our goal is to capture the difference in distribution of these words, among positive and negative reviews. The aim is to obtain context-dependent scores that may replace the polarities coming from SentiWordNet which is a context-independent lexicon [5]. With the help of context-dependent information provided by $\Delta tf * idf$ related features, we expect to better differentiate the positive reviews from negative ones.

We also tried another feature by combining the two information, where we weighted the polarities of all words in the review by their $\Delta tf * idf$ scores (feature F_7).

Table 5. Δtf*idf Features

F_6	$\Delta tf * idf$ scores of all words	$\frac{1}{T}\sum_{j=1..T} \Delta tf * idf(w_j)$
F_7	Weight. avg. pol. of all words	$\frac{1}{T}\sum_{j=1..T} \Delta tf * idf(w_j) \times pol(w_j)$

2.4 Punctuation Features

We have two features related to punctuation. These two features were suggested in [4] and since we have seen that they could be useful for some cases we included them in our sentiment classification system.

Table 6. Punctuation Features

F_8	Number of exclamation marks in the review
F_9	Number of question marks in the review

2.5 Sentence Level Features

Sentence level features are extracted from some specific types of sentences that are identified through a sentence level analysis of the corpus. For instance the first and last lines polarity/purity are features that depend on sentence position; while average polarity of words in subjective/pure etc. sentences are new features that consider only subjective or pure sentences respectively.

Table 7. Sentence-Level Features for a review R

F_{10}	Avg. First Line Polarity	$\frac{1}{N(1)} \sum_{j=1..N(1)} pol(w_{1j})$
F_{11}	Avg. Last Line Polarity	$\frac{1}{N(M)} \sum_{j=1..N(M)} pol(w_{Mj})$
F_{12}	First Line Purity	$\frac{\sum_{j=1..N(1)} pol(w_{1j})}{\sum_{j=1..N(1)} \lvert pol(w_{1j}) \rvert}$
F_{13}	Last Line Purity	$\frac{\sum_{j=1..N(M)} pol(w_{Mj})}{\sum_{j=1..N(M)} \lvert pol(w_{Mj}) \rvert}$
F_{14}	Avg. pol. of subj. sentences	$\frac{1}{\lvert subj(R) \rvert} \sum_{w_j \in subjW(R)} pol(w_j)$
F_{15}	Avg. pol. of pure sentences	$\frac{1}{\lvert pure(R) \rvert} \sum_{w_j \in pure(R)} pol(w_j)$
F_{16}	Avg. pol. of non-irrealis sentences	$\frac{1}{\lvert nonIr(R) \rvert} \sum_{w_j \in nonIr(R)} pol(w_j)$
F_{17}	Δtf*idf weighted polarity of 1st line	$\sum_{j=1..T} \Delta tf * idf(w_{1j}) \times pol(w_{1j})$
F_{18}	Δtf*idf Scores of 1st line	$\sum_{j=1..T} \Delta tf * idf(w_j)$
F_{19}	Number of sentences in review	M

3 Sentence Level Analysis for Review Polarity Detection

We tried three different approaches in obtaining the review polarity. In the first approach, each review is pruned to keep only the sentences that are possibly more useful for sentiment analysis. For pruning, thresholds were set separately for each sentence level feature. Sentences with length of at most 12 words are accepted as short and sentences with absolute purity of at least 0.8 are defined as pure sentences. For subjectivity of the sentences, we adopted the same idea that was mentioned in [23] and applied it on not words, but sentences in this case.

Pruning sentences in this way resulted in lower accuracy in general, due to loss of information. Thus, in the second approach, the polarities in special sentences (pure, subjective, short or no irrealis) were given higher weights while computing the average word polarity. In effect, other sentences were given lower weight, rather than the more severe pruning.

In the final approach that gave the best results, we used the information extracted from sentence level analysis as features used for training our system.

We believe that our main contribution is the introduction and evaluation of sentence-level features; yet other than these, some well-known and commonly used features are integrated to our system, as explained in the next section.

Our approach depends on the existence of a sentiment lexicon that provide information about the semantic orientation of single or multiple terms. Specifically, we use the SentiWordNet [5] where for each term at a specific function, its positive, negative or neutral appraisal strength is indicated (e.g. "good,ADJ, 0.5)

4 Implementation and Experimental Evaluation

In this section, we provide an evaluation of the sentiment analysis features based on word polarities. We use the dominant polarity for each word (the largest polarity among negative, objective or positive categories) obtained from sentiWordNet. We evaluate the newly proposed features and compare their performance to a baseline system. Our baseline system uses two basic features which are the average polarity and purity of the review. These features are previously suggested in [1] and [22] widely used in word polarity-based sentiment analysis. They are defined in Table 3 for completeness. The evaluation procedure we used in our experiments is described in the following subsections.

4.1 Dataset

We evaluated the performance of our system on a sentimental dataset, TripAdvisor that was introduced by [18] and, [19] respectively. The TripAdvisor corpus consists of around 250.000 customer-supplied reviews of 1850 hotels. Each review is associated with a hotel and a star-rating, 1-star (most negative) to 5-star (most positive), chosen by the customer to indicate his evaluation.

We evaluated the performance of our approach on a randomly chosen dataset from TripAdvisor corpus. Our dataset consists of 3000 positive and 3000 negative reviews. After we have chosen 6000 reviews randomly, these reviews were shuffled and split into three groups as train, validation and test sets. Each of these datasets have 1000 positive and 1000 negative reviews.

We computed our features and gave labels to our instances (reviews) according to the customer-given ratings of reviews. If the rating of a review is bigger than 2 then it is labeled as positive, and otherwise as negative. These intermediate files were generated with a Java code on Eclipse and given to WEKA [20] for binary classification.

4.2 Sentiment Classification

Initially, we tried several classifiers that are known to work well for classification purposes. Then, according to their performances we decided to use Support Vector Machines (SVM) and Logistic regression. SVMs are known for being able to handle large feature spaces while simultaneously limiting overfitting, while Logistic Regression is a simple, and commonly used, well-performing classifier. The SVM is trained using a radial basis function kernel as provided by LibSVM [3]. For LibSVM, RBF kernel worked better in comparison to other kernels

on our dataset. Afterwards, we performed grid-search on validation dataset for parameter optimization.

4.3 Experimental Results

In order to evaluate our sentiment classification system, we used binary classification with two classifiers, namely SVMs and Logistic Regression. The reviews with star rating bigger than 2 are positive reviews and the rest are negative reviews in our case, since we focused on binary classification of reviews. Apart from this, we also looked at the importance of the features. The importance of the features will be stated with the feature ranking property of WEKA [20] as well as the gradual accuracy increase, as we add a new feature to the existing subset of features.

For these results, we used grid search on validation set. Then, by these optimum parameters, we trained our system on training set and tested it on testing set.

Table 8. The Effects of Feature Subsets on TripAdvisor Dataset

Feature Subset	Accuracy (SVM)	Accuracy (Logistic)
Basic (F1,F2)	79.20%	79.35%
Basic (F1,F2) + $\Delta TF * IDF$ (F6,F7)	80.50%	80.30%
Basic (F1,F2) + $\Delta TF * IDF$ (F6,F7) + ... Freq. of Subj. Words (F3)	80.80%	80.05%
Basic (F1,F2) + $\Delta TF * IDF$ (F6,F7) + ... Freq. of Subj. Words (F3) + Punctuation (F8,F9)	80.20%	79.90%
Basic (F1,F2) + $\Delta TF * IDF$ (F6,F7) + ... Occur. of Subj. Words (F3-F5)	80.15%	79.00%
All Features (F1-F19)	**80.85**%	**81.45**%

Table 9. Comparative Performance of Sentiment Classification System on TripAdvisor Dataset

Previous Work	Dataset	F-measure	Error Rate
Gindl et al (2010) [6]	1800	0.79	-
Bespalov et al (2011) [2]	96000	-	7.37
Peter et al (2011) [10]	103000	0.82	-
Grabner et al (2012) [7]	1000	0.61	-
Our System (2012)	6000	0.81	-

The results for the best performing feature combinations described in Table 1, are given in Table 8. As can be seen in this table, using sentence level features bring improvements over the best results, albeit small.

4.4 Discussion

As can be seen in the experiments section, our system with the newly proposed features obtains one of the best results obtained so far, except for [2]. Although [2] obtains the best result on a large TripAdvisor dataset, its main drawback is that topic models learned by methods such as LDA requires re-training when a new topic comes. In contrast, our system uses word polarities; therefore it is very simple and fast. For this reason, it is more fair to compare our system with similar systems in the literature.

5 Conclusions and Future Work

In this work, we tried to bridge the gap between word-level polarities and review-level polarity through an intermediate step of sentence level analysis of the reviews. We formulated new features for sentence level sentiment analysis by an in-depth analysis of the sentences. We implemented the proposed features and evaluated them on the TripAdvisor dataset to see the effect of sentence level features on polarity classification. We observed that the sentence level features have an effect on sentiment classification, and therefore, we may conclude that sentences do matter in sentiment analysis and they need to be explored for larger and more diverse datasets such as blogs. For future work, we will evaluate each feature set both in isolation and in groups, and work on improving the accuracy. Furthermore, we will switch to a regression problem for estimating the star rating of reviews.

Sentence level features have other uses since they can be exploited further to identify the essential sentences in the review. We plan to incorporate sentence level features for highlighting the important sentences and review summarization in our open source sentiment analysis system SARE which may be accessed through http://ferrari.sabanciuniv.edu/sare.

Acknowledgements. This work was partially funded by European Commission, FP7, under UBIPOL (Ubiquitous Participation Platform for Policy Making) Project (www.ubipol.eu).

References

1. Ahmed, A., Hsinchun, C., Arab, S.: Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems 26, 1–34 (2008)
2. Bespalov, D., Bai, B., Qi, Y., Shokoufandeh, A.: Sentiment classification based on supervised latent n-gram analysis. In: ACM Conference on Information and Knowledge Management (CIKM) (2011)
3. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines (2001)
4. Denecke, K.: How to assess customer opinions beyond language barriers? In: ICDIM. pp. 430–435. IEEE (2008)

5. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06. pp. 417–422 (2006)
6. Gindl, S., Weichselbraun, A., Scharl, A.: Cross-domain contextualization of sentiment lexicons. Media (2010)
7. Grbner, D., Zanker, M., Fliedl, G., Fuchs, M.: Classification of customer reviews based on sentiment analysis. Social Sciences (2012)
8. Hatzivassiloglou, V., Mckeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics. pp. 174–181. Association for Computational Linguistics (1997)
9. Kim, S.m., Hovy, E., Rey, M.: Automatic detection of opinion bearing words and sentences pp. 61–66
10. Lau, R.Y.K., Lai, C.L., Bruza, P.B., Wong, K.F.: Leveraging web 2.0 data for scalable semi-supervised learning of domain-specific sentiment lexicons. In: Proceedings of the 20th ACM international conference on Information and knowledge management. pp. 2457–2460. CIKM '11, ACM, New York, NY, USA (2011)
11. Mao, Y., Lebanon, G.: Isotonic conditional random fields and local sentiment flow. In: Advances in Neural Information Processing Systems (2007)
12. Martineau, J., Finin, T.: Delta tfidf: An improved feature space for sentiment analysis. In: Adar, E., Hurst, M., Finin, T., Glance, N.S., Nicolov, N., Tseng, B.L. (eds.) ICWSM. The AAAI Press (2009)
13. Mcdonald, R., Hannan, K., Neylon, T., Wells, M., Reynar, J.: Structured models for fine-to-coarse sentiment analysis. Computational Linguistics (2007)
14. Meena, A., Prabhakar, T.V.: Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. Symposium A Quarterly Journal In Modern Foreign Literatures (2), 573–580 (2007)
15. Pang, B., Lee, L.: A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts. Cornell University Library (2004)
16. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of EMNLP. pp. 79–86 (2002)
17. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. 37(2), 267–307
18. The TripAdvisor website. `http://www.tripadvisor.com` (2011), [TripAdvisor LLC]
19. Wang, H., Lu, Y., Zhai, C.: Latent aspect rating analysis on review text data: A rating regression approach. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining pp. 783–792 (2010)
20. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2005)
21. Yu, H.: Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. Proceeding EMNLP 03 Proceedings of the 2003 conference on Empirical methods in natural language processing (2003)
22. Zhai, Z., Liu, B., Xu, H., Jia, P.: Grouping product features using semi-supervised learning with soft-constraints. In: Huang, C.R., Jurafsky, D. (eds.) COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China. pp. 1272–1280. Tsinghua University Press (2010)
23. Zhang, E., Zhang, Y.: Ucsc on rec 2006 blog opinion mining. In: TREC (2006)

24. Zhao, J., Liu, K., Wang, G.: Adding redundant features for crfs-based sentence sentiment classification. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. pp. 117–126 (2008)