

Adaptation and Use of Subjectivity Lexicons for Domain Dependent Sentiment Classification

Rahim Dehkharghani*, Berrin Yanikoglu*, Dilek Tapucu†*, and Yucel Saygin*

* Dept. of Computer Science and Engineering, Sabancı University, Istanbul, Turkey

†Dept. of Computer Engineering, Izmir Institute of Technology, Izmir, Turkey

Email: {rdehkharghani, berrin, dilektapucu, ysaygin}@sabanciuniv.edu

Abstract—Sentiment analysis refers to the automatic extraction of sentiments from a natural language text. We study the effect of subjectivity-based features on sentiment classification on two lexicons and also propose new subjectivity-based features for sentiment classification. The subjectivity-based features we experiment with are based on the average word polarity and the new features that we propose are based on the occurrence of subjective words in review texts. Experimental results on hotel and movie reviews show an overall accuracy of about 84% and 71% in hotel and movie review domains respectively; improving the baseline using just the average word polarities by about 2% points.

Keywords—opinion mining; sentiment analysis; polarity extraction; SentiWordNet; lexicon based methods; machine learning;

I. INTRODUCTION

The amount of information available to individuals is ever-increasing. A considerable part of this information is in textual format, which can be broadly categorized into two main types: facts and opinions [1]. Facts indicate objective information whereas opinions are subjective and indicate the sentiment of the author about an issue.

Opinions can be about anything, e.g. a product such as a movie, a service such as food service of a restaurant or a company such as Amazon. Suppose that the manager of a hotel would like to know the opinion of his guests about their stay. One way is to analyze each review manually which would certainly be time consuming. A more efficient way is to employ a mechanism that could automatically process guest reviews. The goal of sentiment analysis is to facilitate automatic extraction of sentiment from textual data.

In this paper, we address a subproblem of sentiment analysis, namely classifying reviews as either positive or negative. For the classification of reviews, we use subjectivity-based features and experiment with real reviews from the hotel and movie domains. Our contributions could be summarized as: (1) Proposing new subjectivity-based features for sentiment classification (2) Combining domain independent and domain specific subjectivity-based features and evaluating them on two different domains (hotel and movie).

This work was developed in the context of UBIPOL (Ubiquitous Participation Platform for Policy Making) project funded by European Commission, FP7.

The first set of the subjectivity based features that we evaluated is based on word polarities obtained from a domain-independent lexicon and the second one is based on seed word sets adapted to a specific domain. We used the well-known polarity lexicon SentiWordNet, built by Baccianella et. al. [2], for extracting domain-independent subjectivity features. In addition, a domain-specific lexicon of subjective words is used for extracting domain-specific features. To build this lexicon, we selected a subset of the words automatically from the domain-independent subjectivity lexicon built by Hu and Liu [3] that is referred as *InitialSeedWords* in this paper, for two different domains; namely the hotel and movie domains. We refer to this domain-specific lexicon of subjective words as *SubjWords* throughout the paper. The domain-dependent features are based on the occurrence frequencies of *SubjWords* in the reviews. The proposed approach gives a small improvement over the baseline and achieves results compared to other findings in the literature such as [4].

II. RELATED WORK

Our approach to sentiment analysis is a lexicon-based approach, therefore, we mostly report similar works in this section. Lin Pan [5] worked on Chinese language reviews using two sets of positive and negative words, each of which includes more than 4000 words. This work was feature-based and used some predefined templates in sentences. It was applied on different review categories such as hotel review and was able to achieve accuracies higher than 85% in some cases. Graebner et.al. [6] proposed a system that performs the classification of customer reviews of hotels by means of a sentiment analysis. They used a corpus to extract the domain specific lexicon to be used in classification and classified reviews as positive or negative. Taboada et al. [7] took advantage of linguistic resources like dictionaries and built a sentiment analyzer named SO-CAL that was similar to the work done by Polanyi and Zaenen [8].

One of the main drawbacks in lexicon-based approaches is the lack of scalability. For solving this problem, Neviarouskaya and colleagues [9] described methods to automatically generate and score a new sentiment lexicon, called SentiFul, and expanded it through direct synonymy relations and morphologic modifications with known lexical units. They used four types of affixes in their work in sentiment features:

propagating, reversing, intensifying, and weakening. Qiu and colleagues [10] proposed a novel propagation approach that exploits the relations between sentiment words and topics or product features that the sentiment words modify, and also sentiment words and product features themselves to extract new sentiment words.

SentiWordNet is a known resource in sentiment analysis. Ohana and Tierney [4] used the polarity values of words in SentiWordNet to classify movie reviews as positive or negative. In essence, their approach was simple in that they counted the polarity scores of polar words and then improved the approach by adding new features like negation to it. They also used machine learning techniques for classification. This work is explained more in Section V-D because it is the closest one to our approach.

For classifying the product reviews as positive or negative, existing techniques utilize a list of opinion words. Ding and colleagues [11] proposed a holistic lexicon-based approach to increase the accuracy of opinion mining tasks by exploiting external evidences and linguistic conventions of natural language expressions.

Another work similar to ours is proposed by Hamouda and Rohaim [12]. They obtained the polarities of the words inside a document from SentiWordNet and classified reviews as positive or negative based on the summation and average of those polarity scores. They also tried several values as threshold for distinguishing subjective and objective words. Their classification accuracy was around 69% in the best case.

Finally Kaji and Kitsuregawa [13] used structural clues that could extract polar sentences from Japanese HTML documents, and built lexicon from the extracted polar sentences. The key idea was to develop the structural clues. This work was able to provide high precision but not high recall.

III. SUBJECTIVITY-BASED FEATURE EXTRACTION

In supervised training based approaches, sample reviews with known sentiments are used for training a classifier to distinguish between positive and negative reviews, considering the extracted features. Then, given a sample review in testing phase, the same features are extracted and compared to the learned models of positive and negative reviews.

We use the average polarities and weighted polarities of different parts of the review as features, as summarized in Table I. The first five features are computed using word polarities obtained from SentiWordNet, while the last five features are computed using the word polarities obtained from *SubjWords*.

A. Features Based on SentiWordNet (F_1 - F_5)

In SentiWordNet, three scores are assigned to each connotation of a word: positivity, negativity and objectivity [2]. The summation of these three scores equals to one:

$$Pos.Score(w) + Neg.Score(w) + Obj.Score(w) = 1$$

TABLE I
Features extracted for each review

Feature type	Feature name
Domain-independent (Using SentiWordNet)	F_1 : Average polarity of all words
	F_2 : Average polarity of negative words
	F_3 : Average polarity of positive words
	F_4 : Average polarity of last 3 sentences
	F_5 : Average polarity of first 3 sentences
Domain-specific (Using SubjWords)	F_6 : Cumulative frequency of positive words
	F_7 : Cumulative frequency of negative words
	F_8 : Proportion of positive to negative words
	F_9 : Weighted probability of positive words
	F_{10} : Weighted probability of negative words

where w stands for a given word; and the three scores stand for its positivity, negativity and objectivity scores, respectively.

Furthermore, we define the the polarity of a word w as:

$$Pol(w) = Pos.Score(w) - Neg.Score(w)$$

We only consider adjectives and adverbs in a review since they are the most informative terms for sentiment analysis. As a preprocessing step, we eliminated all the words except for the adjectives and adverbs from the reviews. Therefore a word $w_i \in r$ denotes an adjective or an adverb in r . We also do not do word sense disambiguation and use the average polarity of all senses of a word. However, we include all the senses indicated by the POS tag of the word in the context, i.e. if a word is marked as adjective in a sentence, we use only the adjective senses of the word and compute their average over the adjective senses of the word. Then, the average polarity of all words in a review, r , denoted by $AP(r)$ is computed as in (1).

$$AP(r) = \frac{1}{|r|} \sum_{w_i \in r} Pol(w_i) \quad (1)$$

where $|r|$ is the number of words -adjectives and adverbs- in review r and $Pol(w_i)$ is the polarity of the word w_i as defined above.

The first three features (F_1, F_2, F_3) are based on the average polarity concept (AP): F_1 computes the average polarity of all words and F_2 and F_3 compute the average polarity of only the negative and positive words in a review, respectively. A word w is decided as positive if $Pol(w) > 0$, and decided as negative otherwise.

Usually authors express their opinion more directly in first or last parts of a review. In order to factor this information, we used two features (F_4, F_5) as the average polarity of words in last and first three sentences of a review. The features in this section are domain-independent because we extract the polarity of adjectives and adverbs from SentiWordNet which is a domain-independent polarity lexicon.

B. Features Based on SubjWords (F_6 - F_{10})

InitialSeedWords includes 2005 positive, and 4783 negative words, which is filtered to construct a domain dependent set of *SubjWords*. This way we select significantly subjective words for a given domain.

Specifically, we construct the *SubjWords* from the *InitialSeedWords* based on their occurrence in the training set of labeled reviews which we call *AdaptationReviews*. The set of *AdaptationReviews* was also used to calculate the probability distributions for features F_9 and F_{10} to train classifiers. Since evaluation was done using cross-validation on the training set, the *AdaptationReviews* is selected to be a completely different set of reviews to prevent biased testing.

We select a word from the *InitialSeedWords* to be included in *SubjWords* if it appears in the set of *AdaptationReviews*. Our motivation behind this selection is that if a positive word appears in a significant number positive hotel reviews, most probably it will appear among other positive hotel reviews as well. The same argument holds for the negative words. We denote the final selection of positive seed words in *SubjWords* as PS and the final selection of negative seed words in *SubjWords* as NS for the formulation of the subjectivity-based features.

In sentiment analysis, seed word sets are often used by taking into account their occurrences in a review. An alternative is to use measures such as $tf*idf$ (term frequency*inverse document frequency) [14]. The features F_6 and F_7 in our work are based on term frequency values, while F_8 through F_{10} are the newly proposed features based on the occurrence of subjective words in *SubjWords*. Specifically, for F_6 and F_7 we compute the cumulative term frequency of positive and negative seed words for each document in the training set, respectively.

$$F_6(r) = \sum_{t_i \in PS} tf(t_i, r)$$

$$F_7(r) = \sum_{t_i \in NS} tf(t_i, r)$$

Here, $F_6(r)$ is the cumulative frequency of positive seed words in review r ; $tf(t_i, r)$ is the frequency of term t_i in review r . Similarly $F_7(r)$ is the cumulative frequency of negative seed words in review r . Because usually negative reviews are dominated by positive ones and a considerable number of misclassified reviews are negative reviews that have been misclassified as positive, we increased the weight of the negative words by multiplying their frequency by 2; obtaining some improvement in accuracy according to the experimental results.

Since F_6 and F_7 give information about positive or negative term frequencies, we added feature F_8 which is the proportion of positive seed words(the number of occurrences) to the negative ones in a review:

$$F_8(r) = \frac{p + 1}{n + 1}$$

$F_8(r)$ is the proportion of number of positive terms to negative ones in review r ; and p and n are the number of positive and negative seed words, respectively.

Finally features F_9 and F_{10} are the weighted probabilities of positive and negative words in a review, calculated as follows:

$$F_9(r) = p * (1 - P_+(p))$$

$$F_{10}(r) = n * (1 - P_-(n))$$

where $F_9(r)$ is the weighted probability of positive words in a review r ; p is the number of positive seed words in r and $P_+(p)$ is the probability of seeing p positive words in a review. Similarly, $F_{10}(r)$ is the weighted probability of negative words in a review r ; n is the number of negative seed words in the review, and $P_-(n)$ is the probability of seeing n negative words in a review. Probabilities $P_+(p)$ and $P_-(n)$ are calculated from the set of *AdaptationReviews*.

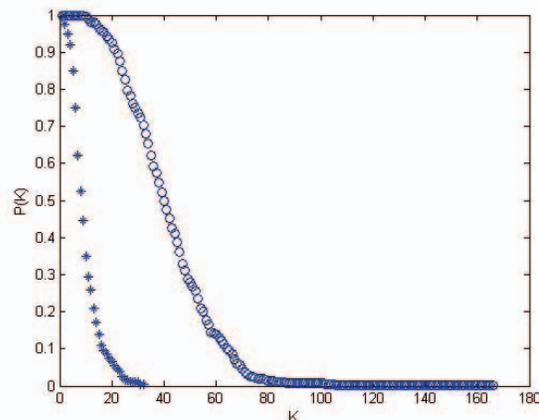


Fig. 1. Plot of $P_+(p)$ as a function of p ('*' represents the hotel and 'o' represents the movie domain).

Note that $(1 - P_+(p))$ increases as $P_+(p)$ decreases; hence, we assign a large weight to an unlikely event such as the occurrence of a very high number of positive words that has a low probability; similarly for $P_-(\cdot)$. Also, while it may seem like $P_+(p)$ and $P_-(n)$ would result in simpler features, the number of positive or negative words, are useful to enhance the feature especially for similar values of $P_+(p)$ (e.g. for $k \in [25 - 35]$ in the hotel domain).

The two probability distributions $P_+(p)$ and $P_-(n)$ are learned over the training set, separately for the two domains. The plot of $P_+(p)$ is displayed for various p values, in Fig. 1. For instance 85% of all hotel reviews contain at least 5 positive words; or correspondingly, $P(5) = 0.85$. Note that the movie reviews are longer, and hence they contain more positive and negative words on average and the probabilities drop more slowly.

IV. SENTIMENT CLASSIFICATION WITH EXTRACTED FEATURES

The classification problem was handled as a binary classification, with 3-star hotel reviews being labelled as objective. Experimental results reported in Section V are obtained using 5-fold cross validation on the training set.

We used three widely used classifiers: Neural Networks, Logistic Regression and Support Vector Machines (SVM) within the Weka 3.6 software. Parameters of these classifiers are as follows. In the SVM classifier, we set γ to 0, loss to 0.1, ϵ to 0.001 and cost to 1.0. In Neural Networks, the learning rate was set to 0.3, the number of hidden layers was set to 1 and the validation threshold was set to 20.

V. EXPERIMENTAL EVALUATION

We implemented and evaluated the subjectivity-based features on real review data sets from two domains, movie and hotel. In this section, we present the experimental results obtained from these data sets.

A. Data Sets

We used the hotel reviews from TripAdvisor corpus [15] and movie reviews from Movie corpus [16]. The TripAdvisor corpus consists of approximately 250000 customer-supplied reviews of 1850 hotels and was provided by [17]. We evaluated the performance of our approach on a randomly chosen subset of the corpus for simplicity. The randomly selected hotel review dataset consists of 6000 reviews half of which are positive, and the other half are negative. We assumed the hotel reviews with rating 1 or 2 as negative, and reviews with 4 or 5 star rating as positive for training. We did not consider the 3 star rating reviews since they do not convey a strong sentiment. We used all reviews in the movie set introduced by [16], including 1000 positive and 1000 negative reviews.

B. Construction of *SubjWords*

The domain independent set of *InitialSeedWords* includes 2005 positive, and 4783 negative words. The set *AdaptationReviews* are randomly selected among the reviews from the hotel and movie domains, and it contains 1000 reviews for hotel and 1000 reviews for movie domains. Consequently we were able to construct the domain dependent set of *SubjWords* consisting of 671 positive and 1393 negative words for the hotel domain and another set of *SubjWords* consisting of 1093 positive and 1977 negative words for the movie domain.

C. Results

We did a comparative evaluation of the subjectivity-based features by evaluating the effectiveness of different feature sets provided in Table I. The results are displayed in Table II, grouped according to basic feature groups in increasing effectiveness and different feature subsets. When the contribution of some particular feature(s) is tested, this is indicated by '+' (e.g. F_6 - F_7 + F_{10}) and positive contributions are highlighted in bold.

We can make several observations regarding the results shown in Table II:

- the best feature group in isolation is based on cumulative term frequencies (F_6 and F_7).
- the accuracy of domain-specific features, F_6 - F_{10} is better than the accuracy of domain-independent ones, F_1 - F_5 .

TABLE II
The accuracy (%) of three classifiers on hotel and movie reviews

Domain	Feature Subset	Accuracy (%)		
		(SVM)	(NN)	(Logistic)
Hotel	Basic: F_1 - F_5	81.58	81.24	81.47
	Pos/Neg. Ratio: F_8	83.37	82.78	82.21
	Weight. Pol.: F_9 , F_{10}	84.45	83.08	82.99
	Cumul. TF: F_6 , F_7	83.56	84.15	83.07
	F_1 - F_5 + F_8	86.36	86.80	86.10
	F_6 - F_7 + F_8	84.52	84.51	83.43
	F_8 + F_9 - F_{10}	85.07	83.48	82.48
	F_6 - F_7 + F_8 - F_{10}	84.50	84.39	83.02
All: F_1 - F_5 + F_6 - F_{10}	87.10	87.08	87.51	
Movie	Basic: F_1 - F_5	62.60	62.00	64.2
	Pos-Neg. Ratio: F_8	67.95	67.50	68.30
	Weight. Pol.: F_9 , F_{10}	69.25	65.85	65.75
	Cumul. TF: F_6 , F_7	70.65	70.25	71.05
	F_1 - F_5 + F_8	69.10	67.50	70.45
	F_6 - F_7 + F_8	67.20	71.25	72.25
	F_8 + F_9 - F_{10}	70.30	70.15	70.80
	F_6 - F_7 + F_8 - F_{10}	68.80	70.95	72.75
All: F_1 - F_5 + F_6 - F_{10}	68.45	71.65	72.85	

- the most useful addition is the positive to negative word ratio (F_8) which is mostly positive.
- in both domains the best results are obtained using all features, except for one experimental setup (the accuracy of the SVM in the movie domain is highest using only F_6 and F_7 , which may be due to suboptimal parameter optimization in SVMs).

Comparing the cumulative polarity (F_6 , F_7) and weighted polarity (F_9 , F_{10}) features, we observe that they both work well with accuracy of around 83% in hotel domain and the accuracy between 65% and 71% in the movie domain. However, F_6 and F_7 features give a little higher accuracy than F_9 and F_{10} . Indeed the *tf* measure on an appropriate set of seed words usually gives good results in opinion mining applications.

The effect of feature F_8 is also good on classification. It alone gave the accuracy of 83% in hotel and 68% in movie domains, which are is similar to the results obtained by features in the domain-specific group. Furthermore, adding F_8 to other feature groups contributed positively. Finally, the set of all features works better than other sets. Hence, we conclude that although most of the features are dependent, each one carries some information that cannot be isolated from the others.

D. Discussion

Some studies in the literature have used SentiWordNet for sentiment classification. The first five features (domain-independent features) can be seen in several works [2] and [12]. However, to the best of our knowledge the domain-specific features, specially F_9 and F_{10} , have not been used before.

Most similar work to ours is [4] where authors worked on movie reviews. What Ohana and Tierney did in their work, is using SentiWordNet for sentiment classification of

TABLE III
Comparison of the accuracy(%) of two approaches

Approach	Our Approach	Our Approach	Ohana [4]
Data Set	(Hotel)	(Movie)	(Movie)
Accuracy	87.51	72.85	69.35

movie reviews. Briefly, their approach is based on extracting some features from SentiWordNet. They used an SVM for classification and their best accuracy is reported as 69.35% (TABLE III). Our approach is similar to theirs but our feature set, specially the second part of features using domain-specific polarity lexicon, is completely different.

VI. CONCLUSION AND FUTURE WORK

In this paper, we worked on subjectivity-based features for sentiment classification. We used two lexicons for feature extraction and experimented on two different domains. We also proposed new subjectivity-based features which improved the sentiment classification accuracy. We used some ready resources: two sets of positive and negative words, used in domain-specific features and also SentiWordNet, used in domain-independent features. The efficiency of domain-specific features was higher than domain-independent ones. After training and testing the system, we achieved an accuracy of about 87% in hotel and 72% in movie domains.

One potential point of improvement of our approach could be in our classification of reviews into subjective and objective. Take for instance: *"this hotel has body building area"*. Although this sentence does not explicitly state any positive or negative connotation, it implies an advantage for the hotel. Having *body building* facility is a positive point for a hotel. Also some idioms are actually subjective but difficult to distinguish: *"staying in this hotel costs an arm and a leg"*; which expresses a negative point for the hotel because it is too expensive.

Our system is completely automatic, however a semi-automatic approach with some human interaction for training the system could give better results.

REFERENCES

- [1] B. Liu, "Sentiment analysis and subjectivity," *Handbook of Natural Language Processing*, vol. 2nd ed, 2010.
- [2] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. of LREC*, 2010.
- [3] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *WWW '05: Proceedings of the 14th international conference on World Wide Web*. New York, NY, USA: ACM, 2005, pp. 342–351.
- [4] B. Ohana and B. Tierney, "Sentiment classification of reviews using SentiWordNet," in *9th. IT & T Conference*, 2009, p. 13.
- [5] H.-Y. Kao and Z.-Y. Lin, "A categorized sentiment analysis of chinese reviews by mining dependency in product features and opinions from blogs," in *Web Intelligence*, J. X. Huang, I. King, V. V. Raghavan, and S. Rueger, Eds. IEEE, 2010, pp. 456–459.
- [6] D. Graebner, M. Zanker, G. Fliedl, and M. Fuchsi, "Classification of customer reviews based on sentiment analysis," in *In 19th Conference on Information and Communication Technologies in Tourism (ENTER)*. Springer, 2012.
- [7] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, Apr. 2011.
- [8] L. Polanyi and A. Zaenen, "Contextual Valence Shifters," in *Computing Attitude and Affect in Text: Theory and Applications*, ser. The Information Retrieval Series, W. B. Croft, J. Shanahan, Y. Qu, and J. Wiebe, Eds. Berlin/Heidelberg: Springer Netherlands, 2006, vol. 20, ch. 1, pp. 1–10.
- [9] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Sentifun: Generating a reliable lexicon for sentiment analysis," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction, IEEE*, 2009.
- [10] G. Qiu, B. L. 0001, J. Bu, and C. Chen, "Expanding domain sentiment lexicon through double propagation," in *IJCAI*, C. Boutilier, Ed., 2009, pp. 1199–1204.
- [11] X. Ding, B. L. 0001, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *WSDM*, M. Najork, A. Z. Broder, and S. Chakrabarti, Eds. ACM, 2008, pp. 231–240.
- [12] A. Hamouda and A. Rohaim, "Reviews classification using sentiwordnet lexicon," in *Journal on Computer Science and Information Technology (OJCSIT)*, Vol. (2), No.(1), 2011.
- [13] N. Kaji and M. Kitsuregawa, "Building lexicon for sentiment analysis from massive collection of HTML documents," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 1075–1083.
- [14] J. Martineau and T. Finin, "Delta tfidf: An improved feature space for sentiment analysis," in *ICWSM*.
- [15] "The TripAdvisor website," <http://www.tripadvisor.com>, 2011, [TripAdvisor LLC].
- [16] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the ACL*, 2004, pp. 271–278.
- [17] H. Wang, Y. Lu, and C. Zhai, "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.