Robust Sense-Based Sentiment Classification

Balamurali A R[1] **Aditya Joshi**[2] **Pushpak Bhattacharyya**[2]
[1] IITB-Monash Research Academy, IIT Bombay
[2]Dept. of Computer Science and Engineering, IIT Bombay
Mumbai, India - 400076
{balamurali,adityaj,pb}@cse.iitb.ac.in

Abstract

The new trend in sentiment classification is to use **semantic features** for representation of documents. We propose a semantic space based on WordNet senses for a supervised document-level sentiment classifier. Not only does this show a better performance for sentiment classification, it also opens opportunities for building a robust sentiment classifier. We examine the possibility of using **similarity metrics** defined on WordNet to address the problem of not finding a sense in the training corpus. Using three popular similarity metrics, we replace unknown synsets in the test set with a *similar* synset from the training set. An improvement of 6.2% is seen with respect to baseline using this approach.

1 Introduction

Sentiment classification is a task under Sentiment Analysis (SA) that deals with automatically tagging text as positive, negative or neutral from the perspective of the speaker/writer with respect to a topic. Thus, a sentiment classifier tags the sentence '*The movie is entertaining and totally worth your money!*' in a movie review as *positive* with respect to the movie. On the other hand, a sentence '*The movie is so boring that I was dozing away through the second half.*' is labeled as *negative*. Finally, '*The movie is directed by Nolan*' is labeled as neutral. For the purpose of this work, we follow the definition of Pang et al. (2002) & Turney (2002) and consider a binary classification task for output labels as positive and negative.

Lexeme-based (bag-of-words) features are commonly used for supervised sentiment classification (Pang and Lee, 2008). In addition to this, there also has been work that identifies the roles of different *parts-of-speech (POS)* like adjectives in sentiment classification (Pang et al., 2002; Whitelaw et al., 2005). Complex features based on parse trees have been explored for modeling high-accuracy polarity classifiers (Matsumoto et al., 2005). Text parsers have also been found to be helpful in modeling valence shifters as features for classification (Kennedy and Inkpen, 2006). In general, the work in the context of supervised SA has focused on (but not limited to) different combinations of bag-of-words-based and syntax-based models.

The focus of this work is to represent a document as a set of sense-based features. We ask the following questions in this context:

1. *Are WordNet senses better features as compared to words?*
2. *Can a sentiment classifier be made robust with respect to features unseen in the training corpus using similarity metrics defined for concepts in WordNet?*

We modify the corpus by Ye et al. (2009) for the purpose of our experiments related to sense-based sentiment classification. To address the first question, we show that the approach that uses senses (either manually annotated or obtained through automatic WSD techniques) as features performs better than the one that uses words as features.

Using senses as features allows us to achieve robustness for sentiment classification by exploiting the definition of concepts (sense) and hierarchical structure of WordNet. Hence to address the second question, we replace a synset not present in the test set with a similar synset from the training set using similarity metrics defined on WordNet. Our results show that replacement of this nature provides a boost to the classification performance.

The road map for the rest of the paper is as follows: Section 2 describes the sense-based features that we use for this work. We explain the similarity-based replacement technique using WordNet synsets

Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pages 132–138,
24 June, 2011, Portland, Oregon, USA

in section 3. Details about our experiments are described in Section 4. In section 5, we present our results and discussions. We contextualize our work with respect to other related works in section 6. Finally, section 7 concludes the paper and points to future work.

2 WordNet Senses as Features

In their original form, documents are said to be in lexical space since they consist of words. When the words are replaced by their corresponding senses, the resultant document is said to be in semantic space.

WordNet 2.1 (Fellbaum, 1998) has been used as the sense repository. Each word/lexeme is mapped to an appropriate synset in WordNet based on its sense and represented using the corresponding synset id of WordNet. Thus, the word *love* is disambiguated and replaced by the identifier *21758160* which consists of a POS category identifier *2* followed by synset offset identifier *1758160*. This paper refers to POS category identifier along with synset offset as synset identifiers or as senses.

2.1 Motivation

We describe three different scenarios to show the need of sense-based analysis for SA. Consider the following sentences as the first scenario.

1. *"Her face **fell** when she heard that she had been fired."*
2. *"The fruit **fell** from the tree."*

The word '*fell*' occurs in different senses in the two sentences. In the first sentence, '*fell*' has the meaning of '*assume a disappointed or sad expression*, whereas in the second sentence, it has the meaning of '*descend in free fall under the influence of gravity*'. A user will infer the negative polarity of the first sentence from the negative sense of '*fell*' in it. This implies that there is at least one sense of the word '*fell*' that carries sentiment and at least one that does not.

In the second scenario, consider the following examples.

1. *"The snake bite proved to be **deadly** for the young boy."*
2. *"Shane Warne is a **deadly** spinner."*

The word *deadly* has senses which carry opposite polarity in the two sentences and these senses assign the polarity to the corresponding sentence. The first sentence is negative while the second sentence is positive.

Finally in the third scenario, consider the following pair of sentences.

1. *"He speaks a **vulgar** language."*
2. *"Now that's real **crude** behavior!"*

The words *vulgar* and *crude* occur as synonyms in the synset that corresponds to the sense '*conspicuously and tastelessly indecent*'. The synonymous nature of words can be identified only if they are looked at as senses and not just words.

As one may observe, the first scenario shows that a word may have *some sentiment-bearing* and *some non-sentiment-bearing senses*. In the second scenario, we show that there may be *different senses of a word that bear sentiments of opposite polarity*. Finally, in the third scenario, we show how *a sense can be manifested using different words*, *i.e.*, words in a synset. The three scenarios motivate the use of semantic space for sentiment prediction.

2.2 Sense versus Lexeme-based Feature Representations

We annotate the words in the corpus with their senses using two sense disambiguation approaches.

As the first approach, **manual sense annotation** of documents is carried out by two annotators on two subsets of the corpus, the details of which are given in Section 4.1. The experiments conducted on this set determine the ideal case scenario- the skyline performance.

As the second approach, a state-of-art algorithm for domain-specific WSD proposed by Khapra et al. (2010) is used to obtain an automatically sense-tagged corpus. This algorithm called **iterative WSD or IWSD** iteratively disambiguates words by ranking the candidate senses based on a scoring function.

The two types of sense-annotated corpus lead us to four feature representations for a document:

1. A group of word senses that have been manually annotated (*M*)

2. A group of word senses that have been annotated by an automatic WSD (*I*)

3. A group of *manually* annotated word senses and words (both separately as features) (*Sense + Words(M)*)

4. A group of *automatically* annotated word senses and words (both separately as features) (*Sense + Words(I)*)

Our first set of experiments compares the four feature representations to find the feature representation with which sentiment classification gives the best performance. Sense + Words(M) and Sense + Words(I) are used to overcome non-coverage of WordNet for some noun synsets.

3 Similarity Metrics and Unknown Synsets

3.1 Synset Replacement Algorithm

Using WordNet senses provides an opportunity to use similarity-based metrics for WordNet to reduce the effect of unknown features. If a synset encountered in a test document is not found in the training corpus, it is replaced by one of the synsets present in the training corpus. The substitute synset is determined on the basis of its similarity with the synset in the test document. The synset that is replaced is referred to as an *unseen synset* as it is not known to the trained model.

For example, consider excerpts of two reviews, the first of which occurs in the training corpus while the second occurs in the test corpus.

1. " *In the night, it is a* ***lovely*** *city and...* "

2. " *The city has many* ***beautiful*** *hot spots for honeymooners.* "

The synset of '*beautiful*' is not present in the training corpus. We evaluate a similarity metric for all synsets in the training corpus with respect to the sense of *beautiful* and find that the sense of *lovely* is closest to it. Hence, the sense of *beautiful* in the test document is replaced by the sense of *lovely* which is present in the training corpus.

The replacement algorithm is described in Algorithm 1. The term *concept* is used in place of *synset* though the two essentially mean the same in this context. The algorithm aims to find a concept *temp_concept* for each concept in the test corpus. The *temp_concept* is the concept closest to some concept in the training corpus based on the similarity metrics. The algorithm follows from the fact that the similarity value for a synset with itself is maximum.

```
Input: Training Corpus, Test Corpus,
Similarity Metric
Output: New Test Corpus
T:= Training Corpus;
X:= Test Corpus;
S:= Similarity metric;
train_concept_list = get_list_concept(T) ;
test_concept_list = get_list_concept(X);
for each concept C in test_concept_list do
    temp_max_similarity = 0 ;
    temp_concept = C ;
    for each concept D in train_concept_list do
        similarity_value = get_similarity_value(C,D,S);
        if (similarity_value > temp_max_similarity) then
            temp_max_similarity= similarity_value;
            temp_concept = D ;
        end
    end
    replace_synset_corpus(C,temp_concept,X);
end
Return X ;
```

Algorithm 1: Synset replacement using similarity metric

The *for* loop over C finds a concept *temp_concept* in the training corpus with the maximum *similarity_value*. The method *replace_synset_corpus* replaces the concept C in the test corpus with *temp_concept* in the test corpus X.

3.2 Similarity Metrics Used

We evaluate the benefit of three similarity metrics, namely LIN's similarity metric, Lesk similarity metric and Leacock and Chodorow (LCH) similarity metric for the synset replacement algorithm stated. These runs generate three variants of the corpus. We compare the benefit of each of these metrics by studying their sentiment classification performance. The metrics can be described as follows:

LIN: The metric by Lin (1998) uses the information content individually possessed by two concepts in addition to that shared by them. The information content shared by two concepts A and B is given by their most specific subsumer (lowest super-

ordinate(*lso*). Thus, this metric defines the similarity between two concepts as

$$sim_{LIN}(A,B) = \frac{2 \times \log Pr(lso(A,B))}{\log Pr(A) + \log Pr(B)} \quad (1)$$

Lesk: Each concept in WordNet is defined through gloss. To compute the Lesk similarity (Banerjee and Pedersen, 2002) between A and B, a scoring function based on the overlap of words in their individual glosses is used.

Leacock and Chodorow (LCH): To measure similarity between two concepts A and B, Leacock and Chodorow (1998) compute the shortest path through hypernymy relation between them under the constraint that there exists such a path. The final value is computed by scaling the path length by the overall taxonomy depth (D).

$$sim_{LCH}(A,B) = -\log\left(\frac{len(A,B)}{2D}\right) \quad (2)$$

4 Experimentation

We describe the variants of the corpus generated and the experiments in this section.

4.1 Data Preparation

We create different variants of the dataset by Ye et al. (2009). This dataset contains 600 positive and 591 negative reviews about seven travel destinations. Each review contains approximately 4-5 sentences with an average number of words per review being 80-85.

To create the manually annotated corpus, two human annotators annotate words in the corpus with senses for two disjoint subsets of the original corpus by Ye et al. (2009). The inter-annotation agreement for a subset(20 positive reviews) of the corpus showed 91% sense overlap. The manually annotated corpus consists of 34508 words with 6004 synsets.

The second variant of the corpus contains word senses obtained from automatic disambiguation using IWSD. The evaluation statistics of the IWSD is shown in Table 1. Table 1 shows that the F-score for noun synsets is high while that for adjective synsets is the lowest among all. The low recall for adjective POS based synsets can be detrimental to classification since adjectives are known to express direct sentiment (Pang et al., 2002).

POS	#Words	P(%)	R(%)	F-Score(%)
Noun	12693	75.54	75.12	75.33
Adverb	4114	71.16	70.90	71.03
Adjective	6194	67.26	66.31	66.78
Verb	11507	68.28	67.97	68.12
Overall	34508	71.12	70.65	70.88

Table 1: Annotation Statistics for IWSD; P- Precision,R-Recall

4.2 Experimental Setup

The experiments are performed using C-SVM (linear kernel with default parameters[1]) available as a part of LibSVM[2] package. We choose to use SVM since it performs the best for sentiment classification (Pang et al., 2002). All results reported are average of five-fold cross-validation accuracies.

To conduct experiments on words as features, we first perform stop-word removal. The words are not stemmed as per observations by (Leopold and Kindermann, 2002). To conduct the experiments based on the synset representation, words in the corpus are annotated with synset identifiers along with POS category identifiers. For automatic sense disambiguation, we used the trained IWSD engine (trained on tourism domain) from Khapra et al. (2010). These synset identifiers along with POS category identifiers are then used as features. For replacement using semantic similarity measures, we used WordNet::Similarity 2.05 package by Pedersen et al. (2004).

To evaluate the result, we use accuracy, F-score, recall and precision as the metrics. Classification accuracy defines the ratio of the number of true instances to the total number of instances. Recall is calculated as a ratio of the true instances found to the total number of false positives and true positives. Precision is defined as the number of true instances divided by number of true positives and false negatives. Positive Precision (PP) and Positive Recall (PR) are precision and recall for positive documents while Negative Precision (NP) and Negative Recall (NR) are precision and recall for negative documents. F-score is the weighted precision-recall

[1] C=0.0,ϵ=0.0010

[2] http://www.csie.ntu.edu.tw/ cjlin/libsvm

Feature Representation	Accuracy	PF	NF	PP	NP	PR	NR
Words	84.90	85.07	84.76	84.95	84.92	85.19	84.60
Sense (M)	89.10	88.22	89.11	91.50	87.07	85.18	91.24
Sense + Words (M)	90.20	89.81	90.43	92.02	88.55	87.71	92.39
Sense (I)	85.48	85.31	85.65	87.17	83.93	83.53	87.46
Sense + Words(I)	86.08	86.28	85.92	85.87	86.38	86.69	85.46

Table 2: Classification Results; M-Manual, I-IWSD, W-Words, PF-Positive F-score(%), NF-Negative F-score (%), PP-Positive Precision (%), NP-Negative Precision (%), PR-Positive Recall (%), NR-Negative Recall (%)

score.

5 Results and Discussions

5.1 *Comparison of various feature representations*

Table 2 shows results of classification for different feature representations. The baseline for our results is the unigram bag-of-words model (Words).

An improvement of 4.2% is observed in the accuracy of sentiment prediction when manually annotated sense-based features (M) are used in place of word-based features (Words). The precision of both the classes using features based on semantic space is also better than one based on lexeme space. Reported results suggest that it is more difficult to detect negative sentiment than positive sentiment (Gindl and Liegl, 2008). However, using sense-based representation, it is important to note that negative recall increases by around 8%.

The combined model of words and manually annotated senses (Sense + Words (M)) gives the best performance with an accuracy of 90.2%. This leads to an improvement of 5.3% over the baseline accuracy [3].

One of the reasons for improved performance is the feature abstraction achieved due to the synset-based features. The dimension of feature vector is reduced by a factor of 82% when the document is represented in synset space. The reduction in dimensionality may also lead to reduction in noise (Cunningham, 2008).

A comparison of accuracy of different sense representations in Table 2 shows that manual disambiguation performs better than using automatic algorithms like IWSD. Although overall classification accuracy improvement of IWSD over baseline is marginal, negative recall also improves. This benefit is despite the fact that evaluation of IWSD engine over manually annotated corpus gave an overall F-score of 71% (refer Table 1). For a WSD engine with a better accuracy, the performance of sense-based SA can be boosted further.

Thus, in terms of feature representation of documents, sense-based features provide a better overall performance as compared to word-based features.

5.2 *Synset replacement using similarity metrics*

Table 3 shows the results of synset replacement experiments performed using similarity metrics defined in section 3. The similarity metric value NA shown in the table indicates that synset replacement is not performed for the specific run of experiment. For this set of experiments, we use the combination of sense and words as features (indicated by *Senses+Words (M)*).

Synset replacement using a similarity metric shows an improvement over using words alone. However, the improvement in classification accuracy is marginal compared to sense-based representation without synset replacement (Similarity Metric=NA).

Replacement using LIN and LCH metrics gives marginally better results compared to the vanilla setting in a manually annotated corpus. The same phenomenon is seen in the case of IWSD based approach[4]. The limited improvement can be due to the fact that since LCH and LIN consider only IS-A

[3]The improvement in results of semantic space is found to be statistically significant over the baseline at 95% confidence level when tested using a paired t-test.

[4]Results based on LCH and LIN similarity metric for automatic sense disambiguation is not statistically significant with α=0.05

Features Representation	SM	A	PF	NF
Words (Baseline)	**NA**	84.90	85.07	84.76
Sense+Words (M)	**NA**	90.20	89.81	90.43
Sense+Words (I)	**NA**	86.08	86.28	85.92
Sense+Words (M)	**LCH**	90.60	90.20	90.85
Sense+Words (M)	**LIN**	90.70	90.26	90.97
Sense+Words (M)	**Lesk**	91.12	90.70	91.38
Sense+Words (I)	**LCH**	85.66	85.85	85.52
Sense+Words (I)	**LIN**	86.16	86.37	86.00
Sense+Words (I)	**Lesk**	86.25	86.41	86.10

Table 3: Similarity Metric Analysis using different similarity metrics with synsets and a combinations of synset and words; SM-Similarity Metric, A-Accuracy, PF-Positive F-score(%), NF-Negative F-score (%)

relationship in WordNet, the replacement happens only for verbs and nouns. This excludes adverb synsets which we have shown to be the best features for a sense-based SA system.

Among all similarity metrics, the best classification accuracy is achieved using Lesk. The system performs with an overall classification accuracy of 91.12%, which is a substantial improvement of 6.2% over baseline. Again, it is only 1% over the vanilla setting that uses combination of synset and words. However, the similarity metric is not sophisticated as LIN or LCH. A good metric which covers all POS categories can provide substantial improvement in the classification accuracy.

6 Related Work

This work deals with studying benefit of a word sense-based feature space to supervised sentiment classification. This work assumes the hypothesis that *word sense is associated with the sentiment* as shown by Wiebe and Mihalcea (2006) through human interannotator agreement.

Akkaya et al. (2009) and Martn-Wanton et al. (2010) study rule-based sentiment classification using word senses where Martn-Wanton et al. (2010) uses a combination of sentiment lexical resources. Instead of a rule-based implementation, our work leverages on benefits of a statistical learning-based methods by using a supervised approach. Rentoumi et al. (2009) suggest an approach to use word senses to detect sentence level polarity using graph-based similarity. While Rentoumi et al. (2009) targets using senses to handle metaphors in sentences, we deal with generating a general-purpose classifier.

Carrillo de Albornoz et al. (2010) create an emotional intensity classifier using affective class concepts as features. By using WordNet synsets as features, we construct feature vectors that map to a larger sense-based space.

Akkaya et al. (2009), Martn-Wanton et al. (2010) and Carrillo de Albornoz et al. (2010) deal with sentiment classification of sentences. On the other hand, we associate sentiment polarity to a document on the whole as opposed to Pang and Lee (2004) which deals with sentiment prediction of subjectivity content only. Carrillo de Albornoz et al. (2010) suggests expansion using WordNet relations which we perform in our experiments.

7 Conclusion & Future Work

We present an empirical study to show that sense-based features work better as compared to word-based features. We show how the performance impact differs for different automatic and manual techniques. We also show the benefit using WordNet based similarity metrics for replacing unknown features in the test set. Our results support the fact that not only does sense space improve the performance of a sentiment classification system but also opens opportunities for building robust sentiment classifiers that can handle unseen synsets.

Incorporation of syntactical information along with semantics can be an interesting area of work. Another line of work is in the context of cross-lingual sentiment analysis. Current solutions are based on machine translation which is very resource-intensive. Using a bi-lingual dictionary which maps WordNet across languages can prove to be an alternative.

References

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity word sense disambiguation. In *Proc. of EMNLP '09*, pages 190–199, Singapore.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proc. of CICLing'02*, pages 136–145, London, UK.

Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervs. 2010. Improving emotional intensity classification using word sense disambiguation. *Special issue: Natural Language Processing and its Applications. Journal on Research in Computing Science*, 46:131–142.

Pdraig Cunningham. 2008. Dimension reduction. In *Machine Learning Techniques for Multimedia*, Cognitive Technologies, pages 91–112.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Stefan Gindl and Johannes Liegl, 2008. *Evaluation of different sentiment detection methods for polarity classification on web-based reviews*, pages 35–43.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Mitesh Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2010. Domain-specific word sense disambiguation combining corpus basedand wordnet based parameters. In *Proc. of GWC'10*, Mumbai, India.

Claudia Leacock and Martin Chodorow. 1998. Combining local context with wordnet similarity for word sense identification. In *WordNet: A Lexical Reference System and its Application*.

Edda Leopold and Jörg Kindermann. 2002. Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46:423–444.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *In Proc. of the 15th International Conference on Machine Learning*, pages 296–304.

Tamara Martn-Wanton, Alexandra Balahur-Dobrescu, Andres Montoyo-Guijarro, and Aurora Pons-Porrata. 2010. Word sense disambiguation in opinion mining: Pros and cons. In *Proc. of CICLing'10*, Madrid,Spain.

Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. 2005. Sentiment classification using word sub-sequences and dependency sub-trees. In *Proc. of PAKDD'05,*, Lecture Notes in Computer Science, pages 301–311.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. of ACL'04*, pages 271–278, Barcelona, Spain.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. volume 10, pages 79–86.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL'04*, pages 38–41.

Vassiliki Rentoumi, George Giannakopoulos, Vangelis Karkaletsis, and George A. Vouros. 2009. Sentiment analysis of figurative language using a word sense disambiguation approach. In *Proc. of the International Conference RANLP'09*, pages 370–375, Borovets, Bulgaria.

Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proc. of ACL'02*, pages 417–424, Philadelphia, US.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proc. of CIKM '05*, pages 625–631, New York, NY, USA.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proc. of COLING-ACL'06*, pages 1065–1072.

Qiang Ye, Ziqiong Zhang, and Rob Law. 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3):6527 – 6535.