Sentiment Classification Using the Meaning of Words

Hadi Amiri and Tat-Seng Chua
Department of Computer Science, National University of Singapore, Singapore, 117417
NUS Graduate School for Integrative Sciences and Engineering, Singapore, 117456
{hadi,chuats}@comp.nus.edu.sg

Abstract
Sentiment Classification (SC) is about assigning a positive, negative or neutral label to a piece of text based on its overall opinion. This paper describes our in-progress work on extracting the meaning of words for SC. In particular, we investigate the utility of sense-level polarity information for SC. We first show that methods based on common classification features are not robust and their performance varies widely across different domains. We then show that sense-level polarity information features can significantly improve the performance of SC. We use datasets in different domains to study the robustness of the designated features. Our preliminary results show that the most common sense of the words result in the most robust results across different domains. In addition our observation shows that the sense-level polarity information is useful for producing a set of high-quality seed words which can be used for further improvement of SC task.

Introduction
The input data for a sentiment analysis system is a set of reviews about an entity such as a person, a product, or a topic. Sentiment classification (SC) is used to discriminate positive and negative reviews about the entities (Pang and Lee, 2008; Liu, 2009). In SC, a set of good representative features are required to determine the polarity of reviews. For example, opinion-bearing words like "amazing" and "terrible" are more important features as they reflect opinion. In this research, we investigate the effectiveness of the sense-level information for SC. More precisely we study the utility and robustness of sense-level features across different domains for SC. Automatic SC could serve as a basis for recommendation of items.

Our preliminary results show that the sense-level information is effective and robust for SC. The results indicate that the use of most common sense of words can improve the accuracy of SC over the baseline in four out of five different domains. However, the commonly-used n-gram features are not robust enough and show widely varying behavior across different domains.

The rest of this paper is organized as follow: Section 2 explains our feature extraction method for SC and Section 3 describes the experimental results and feature analysis. Section 4 reviews the related work, and, finally, Section 5 concludes the paper and explains our future work.

Method
The architecture of our SC system is shown in Figure 1. It has four main components: preprocessing, feature extraction, classification, and SentiWordNet components. We explain each component in the following subsections:

Preprocessing Component
The preprocessing component performs four different tasks. The first is determining the POS tag of the review terms. For POS tagging we used the Stanford Maximum Entropy part of speech tagger (Toutanova and Manning, 2000) because of
its good performance and pre-trained POS tagging models which help us to tag the texts of the reviews without any need for training data. After the POS tagging phase, the sentences and words in each review are tokenized. In this step, we collect the necessary information about the features under studying. The features we consider here are term and part-of-speech n-grams, and review-level features (RF) such as (a) the number of characters, words, and sentences in each review, (b) the vocabulary richness (VR), and (c) information content (IC) scores of the review. These features have been shown to be effective for SC in previous research (Abbasi et al., 2008; Pang and Lee, 2008). We use Simpson’s Reciprocal Index to compute the VR score for reviews (Simpson, 1949). This score is calculated using Equation 1 for a sample review:

\[ VR_r = \frac{N(N-1)}{\sum_{t \in r} t_f(t_f - 1)} \]

where \( N \) is the length of the review and \( t_f \) is the term frequency of the term \( t \) belong to \( r \). The higher value for \( VR_r \) indicates that \( r \) has a richer set of vocabulary. The IC score for each review is computed using Equation 2:

\[ IC_r = -\sum_{t \in r} \log \left( \frac{t_f}{L_r} \cdot \frac{C - DF_t}{C} \right) \]

where \( L_r \) indicates the total number of terms in the review, \( r \), \( C \) is the total number of reviews, \( DF_t \) is the number of reviews that contain the term \( t \), and \( t_f \) is the term frequency of the term \( t \) in \( r \). This score is similar to the the common TF-IDF weighting model in information retrieval.

### Feature Extraction Component

This component generates the associated feature sets for the reviews in each dataset. The features we consider in this research are term unigram (TU), term bigram (TB), POS unigram (PU), and review-level features (RF). We also consider the combination of these features, e.g., TU-PU shows the combination of TU and PU features. This component also extracts features generated based on the sense-level information of the terms. The SentiWordNet component provides the required information for this purpose. These feature sets are SWNOPN, SWNPN, and SWNMCS which will be described in Section 2.4.

### Classification Component

This component is responsible for classifying the documents into positive or negative classes. We use SVM classification for this purpose as it has the best classification performance for sentiment analysis (Pang et al., 2008; Dave et al. 2003) outperforming both the Naive Bayes and Maximum entropy classification methods (Pang et al., 2008).

**Table 1. SentiWordNet information for the word "deep"**

<table>
<thead>
<tr>
<th>Sense</th>
<th>Tag</th>
<th>Pos Score</th>
<th>Neg Score</th>
<th>Term</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>adj</td>
<td>2.87</td>
<td>1.25</td>
<td>deep</td>
<td>+1</td>
</tr>
<tr>
<td>3</td>
<td>noun</td>
<td>0.00</td>
<td>0.00</td>
<td>deep</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>adv</td>
<td>0.25</td>
<td>0.12</td>
<td>deep</td>
<td>+1</td>
</tr>
</tbody>
</table>

**SentiWordNet Component**

SentiWordNet (Esuli and Sebastiani, 2006) is a freely available lexical resource for sentiment analysis in which three polarity scores are assigned to each sense of a word (synset). The polarity scores for each synset \( s \) are three numerical scores \( Obj(s) \), \( Pos(s) \) and \( Neg(s) \) (sum up to one). These scores, respectively, determine the objectivity, positivity, and negativity degrees of the terms in the synset \( s \). To generate the associated scores for each synset, Esuli and Sebastiani (2006) combined the result produced by a committee of eight ternary classifiers. The three polarity scores for a synset are computed based on the proportion of classifiers that have assigned the corresponding label to it. For example, if all the classifiers assign the positive label to a synset, the scores will be \( Obj(s)=0, Neg(s)=0 \) and \( Pos(s)=1 \).

To utilize the SentiWordNet resource for the SC task, we follow the following approaches to construct three feature sets using SentiWordNet sense-level polarity information:

In the first approach, we assign a label from the set \( SO=\{-1,0,+1\} \) to each term \( t \) in review \( r \) using SentiWordNet where "-1" indicates negative polarity, "0" shows no polarity, and "+1" indicates positive semantic orientation. For this purpose, the label of \( t \) will be determined based on the overall polarity of different senses of the term \( t \) in each of its POS tag categories. Here we only consider the noun, adjective, adverb, and verb POS tags. For example, Table 1 shows the labels for the term "deep"; which has 15 adjective, three noun and three verb senses. According to the above scenario, the term "deep" with an adjective POS tag will receive the positive label as the sum of its positive scores (2.87) is greater than the absolute sum of its negative scores (1.25) over all of its fifteen adjective senses. The same is true for its adverb POS tag. However, the term "deep" with a noun POS tag receives a "0" label because its polarity over its three noun senses is zero in SentiWordNet.

We refer to the feature set constructed in this way as SWNOPN. In this feature set, if a term doesn’t occur in the SentiWordNet or if its POS tag is not any of the noun, adjective, adverb and verb word types, then we simply assign it an objective ("0") label. The terms that have similar positive and negative scores in SentiWordNet (totally 2370 terms) are assigned a positive label in SWNOPN. This is because the positive label results in a slightly higher accuracy than the negative label.
We also investigate how the objective terms affect the SC performance. For this purpose, we define the second set of SentiWordNet features, SWNPN, in which all the features with objective label or similar positive and negative scores are removed from SWNOPN feature set. Hence, in SWNPN we assign a label to each term $t$ from the set $SO=\{-1,+1\}$.

The final feature is named SWNMCS in which the most common sense of each term in SentiWordNet is used for labeling the term. So, in this feature set, the biggest polarity score of the most common sense of a term determines the label of the term. For example, if the most common sense of a term has bigger positive polarity score than negative and objective scores, then the term will receive a "+1" label.

**Experimental Results**

In this section, we explain our datasets and SC experiments. All the following experiments are performed based on 10-fold cross validation. We use different datasets from Camera, Camp, Doctor, Music (Whitehead and Yaeger, 2009) and Movie (Pang and Lee, 2004) domains to study the robustness of different feature sets across different domains. Furthermore, we run our experiments using SVMlight package with all parameters set to their default values\(^1\).

**Feature Analysis and Discussion**

We consider TU feature set as the baseline as it has been reported as the best performing features set for SC (Pang and Lee, 2008) and compare the results of using other feature sets against the baseline.

Table 2 shows the results for different feature sets and datasets. A glance of the Table indicates that the accuracy of SC differs greatly from one domain to another. For example, it varies from around 70% in the Music domain to around 88% in the Doctor domain. The high variation in SC performance was observed by some previous works as well (Turney, 2002; Pang and Lee, 2008; Tang et al. 2009). The best performance for each domain is highlighted in Table 2.

The number in bracket indicates the improvement of each corresponding feature set over the baseline (TU). The best accuracies in the Camera, Camp, Doctor, Music and Movie datasets are obtained using SWNPN, SWNMSC, TU, SWNMSC and TUPU feature sets respectively.

Table 2 also shows the robustness of the feature sets over different domains. The column marked with "**" shows the number of times that each feature set improves the accuracy over the baseline (TU). As it is shown, most of the feature sets improve the baseline in one or two domains, but decrease the accuracy in at least three other domains. Such feature sets are not robust enough and can't be considered as a reliable feature set for SC across different domains.

We observed that the only feature sets that show a smooth behavior across the domains is SWNMCS. It improves the accuracy in four out of five domains. For the Doctor domain, though doesn't improve the baseline, it still generates acceptable performance. We believe the robustness of this feature set stems from the fact that the most common sense of the words, as it comes from its name, has higher usage in reviews and therefore using the polarity obtained from the most common sense of the words result in greater and robust SC performance across different domains. The results also indicate the previously studied feature sets like POS tag, bigram, and review-level features are not robust features.

The accuracy of SWNPN is significantly lower than the baseline in most of the domains. However, considering its accuracy and the fact that it skips all the objective terms, we believe that SWNPN produces a set of good seed words that can be used for further improvement of SC accuracy. This is part of our future work to experimentally investigate.

However, we show two short reviews and their SWNPN features for illustration where each term comes with its POS tag and its SentiWordNet score. The terms without

<table>
<thead>
<tr>
<th>Domain &amp; Feature</th>
<th>Camera</th>
<th>Camp</th>
<th>Doctor</th>
<th>Music</th>
<th>Movie *</th>
</tr>
</thead>
<tbody>
<tr>
<td>TU</td>
<td>77.09</td>
<td>88.56</td>
<td>88.31</td>
<td>69.97</td>
<td>85.86</td>
</tr>
<tr>
<td>TU-TB</td>
<td>77.50 (+0.41)</td>
<td>83.65 (-4.91)</td>
<td>86.34 (-1.97)</td>
<td>69.47 (-0.50)</td>
<td>85.71 (-0.15)</td>
</tr>
<tr>
<td>TU-PU</td>
<td>73.96 (+3.13)</td>
<td>85.91 (-2.65)</td>
<td>83.08 (-5.23)</td>
<td>69.69 (-0.28)</td>
<td>87.12 (+1.26)</td>
</tr>
<tr>
<td>TB-PU</td>
<td>70.31 (-6.78)</td>
<td>75.41 (-13.15)</td>
<td>81.46 (-6.85)</td>
<td>69.88 (-0.09)</td>
<td>80.09 (-5.77)</td>
</tr>
<tr>
<td>TU-RF</td>
<td>76.72 (-0.37)</td>
<td>88.67 (+0.11)</td>
<td>87.33 (-0.98)</td>
<td>69.77 (-0.20)</td>
<td>86.62 (+0.76)</td>
</tr>
<tr>
<td>TU-TB-PU</td>
<td>75.61 (+1.48)</td>
<td>83.50 (-5.06)</td>
<td>86.34 (-1.97)</td>
<td>71.24 (+1.27)</td>
<td>85.96 (+0.10)</td>
</tr>
<tr>
<td>TU-RF</td>
<td>75.61 (+1.48)</td>
<td>86.62 (-1.94)</td>
<td>83.74 (-4.57)</td>
<td>70.46 (+0.49)</td>
<td>85.39 (-0.47)</td>
</tr>
<tr>
<td>SWNPN</td>
<td>79.42 (+2.33)</td>
<td>85.76 (-2.80)</td>
<td>84.04 (-4.27)</td>
<td>67.58 (-2.38)</td>
<td>83.17 (-2.68)</td>
</tr>
<tr>
<td>SWNOPN</td>
<td>78.64 (+1.55)</td>
<td>87.68 (-0.88)</td>
<td>87.05 (-1.26)</td>
<td>69.96 (-0.01)</td>
<td>85.70 (+0.16)</td>
</tr>
<tr>
<td>SWNMCS</td>
<td>77.83 (+0.73)</td>
<td>88.88 (+0.32)</td>
<td>87.47 (-0.84)</td>
<td>72.28 (+2.31)</td>
<td>86.54 (+0.68)</td>
</tr>
</tbody>
</table>

Table 2. The average accuracy for polarity detection task over the five datasets

1 SVMlight: http://svmlight.joachims.org/
any score are not in SentiWordNet or do not have any of the noun, adjective, adverb or verb POS tags:

Example 1: This DT doctor NN "0" was VBD most RBS "0" unhelpful JJ "-1" when WRB "0" I PRP called VBD with IN a DT health NN "+1" emergency NN "+1". I PRP was VBD most RBS "0" disappointed JJ "-1".

SWNPN Features: unhelpful JJ "-1" health NN "+1" emergency NN "+1". disappointed JJ "-1".

Example 2: This DT doctor NN "0" does VBZ not RB "-1" have VB "-1" any DT bedside NN "0" manner NN "+1" and CC I PRP would MD not RB "-1" recommend VB "+1" him PRP for IN a DT pediatrician NN "+1". His PRP$ staff NN "0" is VBZ rude JJ "+1" and CC not RB "-1" efficient JJ "+1" in IN any DT manner NN "+1".

SWNPN Features: not RB "-1" have VB "-1" manner NN "+1" recommend VB "+1" pediatrician NN "+1". rude JJ "+1" efficient JJ "+1" manner NN "+1".

It is clear that the SWNPN feature set contains all the important sentiment features of the user reviews. We believe that these features in conjunction with a negation and clause analysis component provide a very good set of seed words for obtaining the polarity of sentences.

Related Work

The research on the interaction between word sense disambiguation and sentiment analysis is quite new. In this area, researchers study the usability of sense level information for different sentiment analysis subtasks such as subjectivity analysis (Akkaya et al., 2009) and disambiguation of sentiment adjectives (Mohtarami et al., 2011).

It has been shown that different senses of a word may have different sentiment orientation. For instance, the word "heavy" in "heavy sleep" produces a positive phrase while in "heavy breathing" indicates a negative phrase. Some researchers take this approach and assign a polarity score to the different senses of the words (Esuli and Sebastiani, 2006; Wiebe and Mihalcea, 2006). Following this approach, we use the sense polarity information of the words to tackle SC at the document level. To the best of our knowledge this research is the first work that utilizes the sense-level polarity information of the words for SC.

Conclusion and Future Work

In this in-progress work, we studied the utility of sense level polarity information for the SC task. For this purpose we used SentiWordNet to construct different feature sets using its sense-level polarity information. We studied the utility of different feature sets and showed that most of the previously investigated feature sets are not robust enough and exhibit varying SC performances across different domains. Our preliminary results show that the sense-level polarity information is useful to improve the overall SC performance. Moreover, we observed that the sense-level polarity information helps to produce a very good set of seed words that can be used for further improve the SC performance. This forms our future work. We also aim to investigate an automatic WSD system to determine the sense of the words for SC in the future work.

References


