Intelligent Techniques for Web Personalization and Recommender Systems
AAAI Technical Report WS-12-09

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.

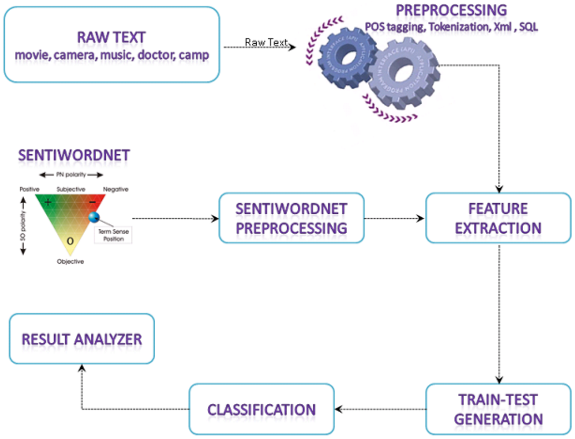

Figure 1. The main components of our system

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 *r*:

$$(1)\, VR_r \quad = \quad \frac{N(N-1)}{\sum_{t\in r} tf_t(tf_t-1)}$$

where N is the length of the review and tf_t 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:

$$(2)\, IC_r \quad = \quad -\sum_{t\in r} Log(\;\frac{tf_t}{L_r} \;*\; \frac{C-DF_t}{C}\;)$$

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 tf_t 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).

#Sense	Tag	Pos Score	Neg Score	Term	Label
15	adj	2.87	1.25	deep	+1
3	noun	0.00	0.00	deep	0
3	adv	0.25	0.12	deep	+1

Table 1. SentiWordNet information for the word "deep"

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 or objective, 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.

Domain & Feature	**Camera**	**Camp**	**Doctor**	**Music**	**Movie**	*
TU	77.09	88.56	**88.31**	69.97	85.86	-
TU-TB	77.50 (+0.41)	83.65 (-4.91)	86.34 (-1.97)	69.47 (-0.50)	85.71 (-0.15)	1
TU-PU	73.96 (-3.13)	85.91 (-2.65)	83.08 (-5.23)	69.69 (-0.28)	**87.12 (+1.26)**	1
TB-PU	70.31 (-6.78)	75.41 (-13.15)	81.46 (-6.85)	69.88 (-0.09)	80.09 (-5.77)	0
TU-RF	76.72 (-0.37)	88.67 (+0.11)	87.33 (-0.98)	69.77 (-0.20)	86.62 (+0.76)	2
TU-TB-PU	75.61 (-1.48)	83.50 (-5.06)	86.34 (-1.97)	71.24 (+1.27)	85.96 (+0.10)	2
TU-PU-RF	75.61 (-1.48)	86.62 (-1.94)	83.74 (-4.57)	70.46 (+0.49)	85.39 (-0.47)	1
SWNPN	**79.42 (+2.33)**	85.76 (-2.80)	84.04 (-4.27)	67.58 (-2.38)	83.17 (-2.68)	1
SWNOPN	78.64 (+1.55)	87.68 (-0.88)	87.05 (-1.26)	69.96 (-0.01)	85.70 (-0.16)	1
SWNMCS	77.83 (+0.73)	**88.88 (+0.32)**	87.47 (-0.84)	**72.28 (+2.31)**	86.54 (+0.68)	**4**

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

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 SVM^{light} 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

[1] SVM^{light}: 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

Abbasi A., Chen H., and Salem A. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transaction on Information System 26(3).

Akkaya C., Wiebe J., and Mihalcea R. 2009. Subjectivity Word Sense Disambiguation. In Proceedings of EMNLP.

Dave K,, Lawrence S,, and Pen-nock D. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of WWW.

Esuli A., and Sebastiani F. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In Proceedings of LREC-06.

Liu B. 2009. Web Data Mining: Exploring Hyper-links, Contents, and Usage Data. Springer, 2nd edition.

Mohtarami M., Amiri H., Lan M., Tan C. 2011. Predicting the Uncertainty of Sentiment Adjectives in Indirect Answers. In Proceedings of CIKM '11.

Pang B., and Lee L. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, Proceedings of ACL.

Pang B., and Lee L. 2008. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1).

Tang H., Tan S., and Cheng X. 2009. A survey on sentiment detection of reviews. In proceeding of Expert Systems with Applications 36,

Toutanova K., and Manning C. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of EMNLP/VLC-2000.

Turney P. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of ACL.

Whitehead M., and Yaeger L. 2009. Building a General Purpose Cross-Domain Sentiment Mining Model. In proceeding of CSIE.

Wiebe J., and Mihalcea R. 2006. Word sense and subjectivity. In Proceedings of COLING/ACL.