Semantic Smoothing for Twitter Sentiment Analysis

Hassan Saif, Yulan He and Harith Alani

Knowledge Media Institute, The Open University, United Kingdom
{h.saif,y.he,h.alani}@open.ac.uk

Abstract. Twitter has brought much attention recently as a hot research topic in the domain of sentiment analysis. Training sentiment classifiers from tweets data often faces the data sparsity problem partly due to the large variety of short forms introduced to tweets because of the 140-character limit. In this work we propose using semantic smoothing to alleviate the data sparseness problem. Our approach extracts semantically hidden concepts from the training documents and then incorporates these concepts as additional features for classifier training. We tested our approach using two different methods. One is shallow semantic smoothing where words are replaced with their corresponding semantic concepts; another is to interpolate the original unigram language model in the Naive Bayes (NB) classifier with the generative model of words given semantic concepts. Preliminary results show that with shallow semantic smoothing the vocabulary size has been reduced by 20%. Moreover, the interpolation method improves upon shallow semantic smoothing by over 5% in sentiment classification and slightly outperforms NB trained on unigrams only without semantic smoothing.

1 Introduction

Few years after the explosion of Web 2.0, microblogs and social networks are now considered as the most popular forms of communication. Through a platform like Twitter, tones of information, which reflects people's opinions and attitudes, are published and shared among users everyday. By monitoring and analysing opinions, organisations can detect the level of satisfaction or intensity of complaints about certain products and services; policy makers and politicians are able to analysis the public opinions about their policies or political issues.

Previous work on twitter sentiment analysis [1–3] addresses the problem as a text classification task where classifiers are built using various machine learning methods and trained on labeled corpora where each tweet is labeled as positive or negative using features like unigrams, bigrams, part of speech tags, etc. Although this line of work shows relatively good classification results, the typical use of classification methods like Naive Bayes on tweets data usually poses several challenges. One major challenge is the data sparsity problem partly due to a large variety of short forms found at tweets content due to the 140-character limit.

In this work we propose our semantic smoothing approach to alleviate the data

sparseness problem. The main idea is to extract semantically hidden concepts from data and then incorporate these concepts as additional features for classifier training. We investigate two different ways to realise semantic smoothing. One is called shallow semantic smoothing which simply replaces words in the training dataset with their corresponding semantic concepts. Another is called the interpolation method where we interpolate the original unigram language model in the Naive Bayes (NB) classifier with the generative model of words given semantic concepts. Our experimental results on the Twitter sentiment dataset show that using shallow semantic smoothing, the sentiment classification accuracy drops by nearly 5% compared to NB trained on unigrams solely. However, such accuracy loss is recovered using the interpolation method with the final result being slightly better than that obtained by NB without semantic smoothing. Although the improvement is only marginal, our preliminary experimental results give promising directions for future work since there is still room for improvement, for example, feature selection and tuning of the interpolation coefficient.

2 Approach

Shallow Semantic Smoothing One technique to solve the data sparsity problem is what we may call shallow semantic smoothing where we extract semantically hidden concepts from data and then replace these concepts with their related entities. For example, the sentence *"downloading apps for my iPhone! So much fun :)"* contains the entity *"iPhone"* which will be replaced by the concept *"Product"*. Other entities appeared in document corpora which are semantically similar to the entity *"iPhone"* will also be replaced by the concept *"Product"*. This approach should reduce the vocabulary size of the training data, which in turn makes the data less sparse.

Semantic Smoothing for Naive Bayes Classifier Naive Bayes (NB) is a supervised probabilistic learning method. It is widely used for sentiment classification. The assignment of class c to a given document d can be computed as:

$$C(d) = \arg\max_{c \in C} P(c|d) = \arg\max_{c \in C} P(c) \prod_{1 \leq i \leq n_d} P(w_i|c) \tag{1}$$

Where $P(c)$ is the class prior and $P(w_i|c)$ is the conditional probability of word w_i occurring in a document of class c. There are several types of NB classifier such as binomial and multinomial NB models. In this study we conducted all experiments using the multinomial model. A well-known implementation of the multinomial class model is the maximum likelihood estimate with Laplace smoothing:

$$P(w|c) = \frac{N(w,c) + 1}{\sum_{w' \in V} N(w'|c) + |V|} \tag{2}$$

Although using Laplace smoothing helps to prevent zero probabilities of the "hidden" words, it assigns equal prior probabilities to all of these words.

We could however interpolate the unigram language model in NB with the generative model of words given semantic concepts. Thus, the new class model with

semantic smoothing has the following formula:

$$P_s(w|c) = (1-\alpha)P_u(w|c) + \alpha \sum_j P(w|s_j)P(s_j|c) \quad (3)$$

Where $P_s(w|c)$ is the unigram class model with semantic smoothing, $P_u(w|c)$ is the unigram class model with maximum likelihood estimate, s_j is the j-th concept of the word w, $P(s_j|c)$ is the distribution of semantic concepts in training data of a given class and it can computed via the maximum likelihood estimation. $P(w|s_j)$ is the distribution of words in the training data given a concept and it can estimated by the EM algorithm. Finally, the coefficient α is used to control the influence of the semantic mapping in the new class model.

3 Experiments

Twitter Sentiment[1] dataset was built based on the work of Go et al. [2]. The training dataset consists of 1.6 million tweets. It was collected between the 6th of April and the 25th of June 2009 with equal number of positive and negative tweets labeled based on emoticons appeared in the tweets. The testing dataset was manually collected and annotated. It consists of 177 negative and 182 positive tweets. We selected a balanced subset of 60,000 tweets from the Twitter Sentiment corpus for training and the same test set for testing. Stop words were removed and all words were stemmed.

Table 1. Top 5 concepts with the number of their associated entities.

Concept	Person	Company	City	Country	Organisation
# of Entities	4954	2815	1575	961	614

We used three different third-party services to extract entities from both training and testing dataset, Zemanta, OpenCalais, and AlchemyAPI.[2] A quick and manual comparison of a randomly selected 100 extracted entities with their corresponding semantic concepts showed that AlchemyAPI performs better than the others on tweet messages in terms of the quality and the quantity of the extracted concepts. Using AlchemyAPI on our training dataset, we extracted a total of 15,139 entities, which are mapped to 30 distinct concepts. Table 1 shows the top five concepts with the number of entities associated to them.
We conducted a set of experiments using both shallow semantic smoothing and the interpolation method for Naive Bayes to evaluate the sentiment classification accuracy. For the interpolation method, we set the interpolation coefficient α to 0.8 without any tuning. It can be observed from Table 2 that NB trained on unigrams only gives a sentiment classification accuracy of 81%. However, with

[1] http://twittersentiment.appspot.com/
[2] Zemanta(http://www.zemanta.com/); OpenCalais(http://www.opencalais.com/); AlchemyAPI(http://www.alchemyapi.com/)

shallow semantic smoothing, the accuracy drops by nearly 5% due to information loss caused by the mere use of semantic concept replacement. This can be improved by performing a selective statistical replacement which is determined based on the contribution of each concept towards making a better classification decision. Indeed, using the interpolation method, the accuracy loss is recovered. Although the improvement over NB without semantic smoothing is only marginal, the results can be enhanced further by, for example, 1) performing feature selection before training the classifiers; 2) better controlling the influence of semantic smoothing in the class model by adaptively setting the interpolation coefficient α; 3) improving the quality of the entity-concept extraction method.

Table 2. Sentiment classification results using different method. The original method is NB trained on unigrams only.

Method	Accuracy
No semantic smoothing	81.0%
Shallow Semantic Smoothing	76.3%
The interpolation method	81.3%

4 Conclusion and Future Work

We have proposed two semantic smoothing methods to address the data sparsity problem. The first one, shallow semantic smoothing, replaces entities in tweet messages with their related concepts. It successfully reduces the vocabulary size by nearly 20%. However, following the naive method by simply replacing all entities with their concepts leads to loss of information, which in turn affects the sentiment classification results as shown in Table 2. The second one, the interpolation method for NB, semantically smooths the unigram language model by better incorporating concepts in the classification process. Preliminary results show that the interpolation method outperforms shallow semantic smoothing and attains slightly better results than NB without semantic smoothing. Our preliminary experimental results give promising directions for future work since there is still room for improvement such as feature selection and adaptive tuning of the interpolation coefficient so as to better control the influence of semantic smoothing.

Acknowledgement This work is funded by the EU project ROBUST (grant number 257859).

References

1. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of COLING. pp. 36–44 (2010)
2. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (2009)
3. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC 2010 (2010)