Topic-Dependent Sentiment Classification on Twitter

Steven Van Canneyt, Nathan Claeys, and Bart Dhoedt

Department of Information Technology, Ghent University - iMinds, Belgium
{steven.vancanneyt,nathan.claeys,bart.dhoedt}@ugent.be

Abstract. In this paper, we investigate how discovering the topic dicussed in a tweet can be used to improve its sentiment classification. In particular, a classifier is introduced consisting of a topic-specific classifier, which is only trained on tweets of the same topic of the given tweet, and a generic classifier, which is trained on all the tweets in the training set. The set of considered topics is obtained by clustering the hashtags that occur in the training set. A classifier is then used to estimate the topic of a previously unseen tweet. Experimental results based on a public Twitter dataset show that considering topic-specific sentiment classifiers indeed leads to an improvement.

1 Introduction

Twitter is an excellent source of opinions, as it gives us access to the unprompted views of a broad set of users on particular products or events. The opinions or expressions of sentiment about organizations, products, events and people has proven extremely useful for marketing [8] and social studies [13]. Often it is especially important to quickly detect negative opinions, so a company can respond to any criticism in a timely manner. Therefore, we will focus on detecting the tweets expressing negative sentiments.

The sentiment of words used in a tweet are often dependent on the topic of that tweet. For example the tweet 'So I juuuust started the first amazing 15 minutes of The Last of Us, when my ps3 shuts off and the red light started blinking' with a negative sentiment label contains the word 'amazing' which in general indicates a positive sentiment. However as this tweet is situated in the 'Game console' topic, 'red' is associated with the crash of the ps3 which always show the infamous red light blinking. Therefore, we propose a methodology that directly uses the topics of tweets to improve the sentiment classification. We consider a cluster of similar hashtags as a topic. For each cluster we train two classifiers: one classifier aimed at recognising tweets that talk about the corresponding topic, and one classifier aimed at detecting negative opinions in tweets talking about this topic. Given a previously unseen tweet, we use the classifiers of the former type to determine the most likely topic. Then we use the corresponding topic-specific sentiment classifier to estimate the sentiment of the tweet.

A. Hanbury et al. (Eds.): ECIR 2015, LNCS 9022, pp. 441–446, 2015.

The remainder of this paper is structured as follows. We start with a review of related work in Section 2. Next, in Section 3, we describe our topic-dependent classifier. Section 4 explains how the topics of the tweets are estimated. Details on the considered training data, test data and preprocessing steps are provided in Section 5. Subsequently, Section 6 presents the experimental results. Finally, we conclude our work and discuss future work in Section 7.

2 Related Work

Early work on sentiment analysis focused largely on blogs and reviews. Das et al. [4], for instance, used lexical resources to decide whether a post on a stock message board expresses a positive or negative sentiment by the presence of sentiment words. In addition, linguistic rules were used to deal with e.g. negation in sentences. The authors of [11] researched the performance of various machine learning based classifiers for sentiment classification of movie reviews. A more comprehensive survey about sentiment analysis on documents such as reviews can be found in [10].

In recent years, sentiment classification in Twitter has gained a lot of attention. This introduced additional challenges as tweets tend to be very short and noisy compared to reviews and blogs. The methodology described in this paper is based on the machine learning technique introduced by Go et al. [6]. They tested the suitability for sentiment classification of a number of standard classifiers, including Naive Bayes, SVM and Maximum Entropy classifiers. These classifiers were trained using emoticons in the tweets as labels, together with different types of features for the text of the tweets such as unigrams, bigrams and part-of-speech (POS) features. As using bigrams and POS features in addition to unigrams did not increase the performance of the classifiers, we only consider unigrams in this paper. The research of Bifet et al. [2] notes that the accuracy of the sentiment classifiers needs to be nuanced as it is shown that these classification algorithms often favour the most common class. This typically results in good classification performance for this class at the cost of the smaller classes. By focussing on detecting negative tweets we avoid this problem.

The hashtags used in Twitter have been used by several in the context of sentiment analysis. In addition to using hashtags as unigram features [6], they have been used as sentiment labels [5]. In the paper of Davidov et al. [5], the hashtags #happy, #sad, #crazy and #bored were used to label the training data of a classifier. Similar to our approach, Wang et al. [14] considered hashtags as topics. However, their objective is to estimate the sentiment related to a hashtag. In contrast, we consider hashtags clusters as topics and use topic-specific classifiers to improve the quality of the sentiment detection for individual tweets.

3 Sentiment Classification

The sentiment classifier estimates the probability that a tweet is negative, which allows us to sort the tweets according to the likelihood that a tweet is negative.

The sentiment classifier consists of one generic classifier C^K and a topic-specific classifier C^d. The generic classifier C^K is trained on all the tweets in training set K and estimates the generic probability that a tweet t contains negative sentiment, i.e. $P_t(\text{neg}|K)$. The topic-specific classifier C^d is only trained on the tweets in K of topic d. The classifier C^{d_t} estimates the topic-specific probability that a tweet t of topic d_t is negative, i.e. $P_t(\text{neg}|d_t)$.

For the generic and topic-specific classifiers, the Naive Bayes Multinomial classifier [9] implementation of MOA [3] is used. The feature vector V_t of the tweet t, which is used as input for these classifiers, is constructed using a bag-of-words approach. The components of vector V_t are associated with a word that appears in dictionary W. This dictionary W is the set of all words occuring in the tweets of training set K. For feature vector V_t of tweet t, the component comp_w associated with word $w \in W$ is given by:

$$\text{comp}_w = \begin{cases} \frac{\max(p_w, n_w)}{p_w + n_w} \times \frac{|K|}{p_w + n_w} & \text{if } w \in t \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

with p_w and n_w being the absolute frequency of occurrences of word w in respectively positive or negative tweets in K. The first part of this equation ensures that words which occur often in only one sentiment category (positive or negative) have higher associated component values. The second part ensures that words which occur in a lot of tweets of K have lower associated component values. We also experimented with binary features and term frequency features, but as initial experiments showed that these alternatives yield worse results, we will not consider them in the remainder of the paper.

We finally define the probability $P_t(\text{neg})$ that a tweet t is negative as follows:

$$P_t(\text{neg}) = \lambda \cdot P_t(\text{neg}|d_t) + (1 - \lambda) \cdot P_t(\text{neg}|K) \tag{2}$$

with d_t the topic of tweet t, and $\lambda \in [0, 1]$.

4 Topic Classification

The definition of $P_t(\text{neg})$ in (2) assumes that we already know the topic of a tweet. Therefore, a topic classification algorithm is used to classify each tweet into a fitting topic. In this paper, topics are defined by the hashtag clusters that are present in the collection of tweets K. First, the hashtags are clustered into topics D using the Spectral Clustering algorithm with the cut-off threshold of τ [7,12]. The co-occurence distance between two hashtags $h1$ and $h2$ is used as distance measure:

$$\text{distance}(h1, h2) = 1 - \left(\frac{n_{h1,h2}}{\sum_{i=1}^{|H|} n_{h1,hi}} + \frac{n_{h1,h2}}{\sum_{i=1}^{|H|} n_{h2,hi}}\right) \times \frac{1}{2} \tag{3}$$

with H the set of hashtags that occur in the tweets of training set K, and $n_{h1,h2}$ the number of times hashtag $h1$ and $h2$ occur together in the tweets of K. The idea of using this distance measure is that hashtags which co-occur in the same tweets are associated with a similar or even the same topic such as '#cod' and '#callofduty'. As a result of this step, we have a number of clusters of hashtags. We interpret each of these clusters as a topic. The set K_D contains all tweets of K that have at least one hashtag associated with a cluster. Second, tweets in K_D are associated to their corresponding topic. Third, the binary bag-of-words feature vectors of the tweets in K_D are used to train a Naive Bayes Multinomial classifier [3,9], whereby the topics of the tweets are used as labels. Finally, this classifier is used to estimate the topics of the tweets in $K \setminus K_D$ and U. This topic classification approach is based on the methodology described in [1].

5 Data Collection and Preprocessing

We use the public available Stanford Twitter Sentiment corpus[1] introduced by Go et al. [6]. They obtained training set K by automatically labeling tweets based on their emoticons. The use of emoticons as noisy labels makes it easy to extract a large set of training data. In particular, the Twitter API was first queried between April 6, 2009 and June 25, 2009 using query ':(' and ':)' to extract tweets with respectively negative and positive sentiment. Second, the emoticons in the tweets were stripped off and retweets were removed. Finally, the first 800 000 tweets with positive emoticons and the first 800 000 tweets with negative emoticons were considered as training set K. The test set U constructed by [6] contains tweets collected by querying the Twitter API with queries indicating products, companies and people. The obtained tweets were manually annotated resulting in 177 negative, 182 positive and 139 neutral tweets.

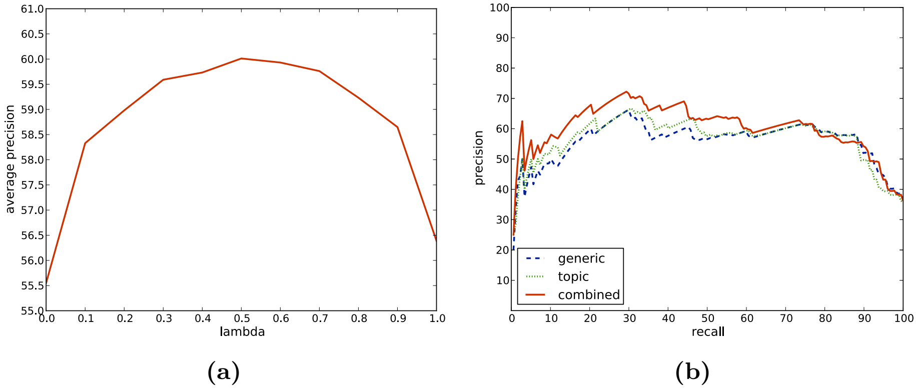
Similar as described in [6], all collected tweets were preprocessed to reduce the feature space. In particular, the words of the tweets were converted to lower case and Porter stemmed, and user mentions and URLs were replaced by respectively 'USER_TOKEN' and 'URL_TOKEN'.

6 Results

To evaluate the advantage of using topic-specific classifiers, we compare the result of the proposed classifier with the result of using the generic classifier alone, i.e. $P_t(\text{neg}|K)$. We also evaluate the performance of using the topic-specific classifiers without the generic classifier, i.e. $P_t(\text{neg}|d_t)$. The classifiers are used to estimate the probability that the tweets in the test set are negative, and to rank them based on their associated probability. To evaluate the quality of the ranking, the average precision metric (AP) is used. We empirically set the cut-off threshold τ for the Spectral Clustering algorithm to 0.98.

The average precision for different λ values for equation (2) are shown in Figure 1(a) (for test set U). Note that only the generic classifier $P_t(\text{neg}|K)$ is used when $\lambda = 0$, and only the topic-specific classifier $P_t(\text{neg}|d_t)$ is used when

[1] http://help.sentiment140.com/



(a)

(b)

Fig. 1. (a) Average precisions for different λ values. (b) Precision-recall curves of the generic classifier (baseline), the topic-specific classifier and the combined classifier.

$\lambda = 1$. As can be seen, the curve is more or less symmetric with an maximum average precision when $\lambda = 0.5$ is used. The average precision of the combined classifier with optimal λ (AP = 60.01%) is 4.5 percentage points higher than when the generic classifier (AP = 55.55%) is used. To determine if the difference in quality of the classifications are statistically significant when the different approaches are used, we consider the sign test on the classification accuracy metric. In particular, we obtained a classification accuracy of 82.3% when the combined classifier with $\lambda = 0.5$ is used, which is statistically significant better then when the generic classifier (accuracy of 79.9%) is used (sign test, $p < 0.01$). Finally, the precision-recall curves of the combined classifier with $\lambda = 0.5$, the generic classifier and topic-specific classifier are shown in Figure 1(b).

The following is an example tweet where the topic classifier shows a better probability than the generic classifier: ‘I still love my Kindle2 but reading The New York Times on it does not feel natural’. This tweet contains a negative label, however the generic classifier classifies this as positive. This is most likely because the word ‘love’ is the only generic word which gives a real idea about the sentiment. The topic classifier however sees the word ‘natural’ as negative, while the generic classifier does not. This can be explained because in the cluster ‘...#amazon #book #kindle...’ the word ‘natural’ refers to the problem that some users did not find reading on the Kindle2 as natural as reading a book or a newspaper. This is an example of a topic-specific feature that has a strong meaning in the topic that is non-existent in the general tweet corpus because the feature is widely used in general tweets. This sort of features allow the topic-specific classifier to make corrections to the negative probability of the generic classifier.

7 Conclusions and Future Work

We proposed a methodology to rank tweets based on the probability that they express negative sentiment. To this end, we have interpolated a generic language model for negative sentiment and a topic-specific model. In this way we can take advantage of the robustness of a generic classifier, which can be trained on a much larger training set, with the ability of topic-specific classifiers to pick up on context-specific expressions of sentiment. We used a fixed set of topics based on the hashtags from the tweets in the training set. As the topics that are discussed in tweets change over time, in future work we will consider a topic detection approach which evolves over time.

Acknowledgments. Steven Van Canneyt is funded by a Ph.D. grant of the Agency for Innovation by Science and Technology (IWT).

References

1. Antenucci, D., Handy, G., Modi, A., Tinkerhess, M.: Classification of tweets via clustering of hashtags. In: EECS 545 Project, pp. 1–11 (2011)
2. Bifet, A., Frank, E.: Sentiment knowledge discovery in Twitter streaming data. In: Pfahringer, B., Holmes, G., Hoffmann, A. (eds.) DS 2010. LNCS (LNAI), vol. 6332, pp. 1–15. Springer, Heidelberg (2010)
3. Bifet, A., Holmes, G., Kirkby, R., Pfahringer, B.: MOA: Massive Online Analysis. Journal of Machine Learning Research 11, 1601–1604 (2010)
4. Das, S., Chen, M.: Yahoo! for Amazon: Extracting market sentiment from stock message boards. Management Science 53(9), 1375–1388 (2007)
5. Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using Twitter hashtags and smileys. In: Proc. of the 23rd Int. Conf. on Computational Linguistics (2010)
6. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. In: CS224N Project Report, Stanford (2009)
7. Hall, M., National, H., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Explorations 11(1) (2009)
8. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent Twitter sentiment classification. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 151–160 (2011)
9. Mccallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. In: Proc. of the AAAI-98 Workshop on Learning for Text Categorization, pp. 41–48 (1998)
10. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Inf. Retrieval 2(1-2), 1–135 (2008)
11. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing, pp. 79–86 (May 2002)
12. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Pattern Analysis and Machine Intelligence 22(8), 888–905 (2000)
13. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment in Twitter events. Journal of the American Society for Inf. Science and Technology 62(2), 406–418 (2011)
14. Wang, X., Wei, F., Liu, X., Zhou, M., Zhang, M.: Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach. In: Proc. of the 20th ACM Int. Conf. on Inf., pp. 1031–1040 (2011)