Twitter-based Sentiment Analysis

Jakob Gruber
0203440

Matthias Krug
0828965

Stefanie Plieschnegger
0926102

Christian Proske
1328245

Mino Sharkhawy
1025887

ABSTRACT
Sentiment analysis has become very popular in recent years and especially Twitter provides a lot of data to a huge amount of topics which can be processed and classified to provide an overall opinion. However, classification of Twitter-based data is somehow different to traditional text mining and introduce some additional challenges. In this paper typical problems are discussed that go along with classification of tweets. We will also shortly discuss two popular machine learning algorithms (Naive Bayes and SVM) for sentiment analysis and explain how a classifiers are evaluated.

Categories and Subject Descriptors
H.3 [**Information Systems**]: Information Search and Retrieval

General Terms
Theory

Keywords
Sentiment Analysis, Opinion Mining, Twitter, Classifier

1. INTRODUCTION
The general opinion about a specific product or service has certainly a great influence on its reputation. People often want to know what others think about a special product they are willing to buy, about a new movie, or about a hotel they are going to book. But also companies may interested in its customers' opinions, politicians may wish to receive feedback, or social organizations may have interest in an ongoing debate. [3]. The world wide web provides many ways for people to distribute their experiences and sentiments. Machine learning algorithms make it easier to process and evaluate those sentiments and are therefore able to provide an overall opinion to a certain topic. This kind of analyzing is called sentiment analysis or opinion mining [2, 4]. Clearly, there are some challenges when assessing the opinion of people, especially when classifying microblogging services like Twitter [1]. The underlying paper gives an overview about different sentiment analysis approaches and outlines special problems related to the classification of microblogging services. These challenges are discussed in section 2. Section 3 gives some basic information about the Twitter API. In section 4 and 5 it is describes approaches of preprocessing and classifying tweets. In section 6 some related work is introduced and section 7 concludes the paper.

2. ANALYZING DATA
The analyzing process of text and the classification whether its content is rather positive, negative or may be considered as neutral, is the core functionality in sentiment analysis. We are going to discuss general considerations of classifying data first of all, before outlining additional challenges related to Twitter-based data.

2.1 Text Mining vs. Sentiment Analysis
One would assume, that text mining and sentiment analysis are very similar. Text mining e.g. may deal with classifying documents by topic, which represents one of the easier tasks. Topic-based text classification tries to match a text into a category like sport, politics etc. and therefore topic-related words are identified to classify a text. However, sentiment analysis requires to focus on typical "sentiment words" for example hate, love, like, regret. When it comes to identify the overall sentiment of a text it turns out that it may contains several aspects (e.g. negative and positiver remarks) or may not even contain any "signal words" (e.g. terrible, awful, bad) but still has a negative meaning. To underline this problem, here are some examples:

"My new smartphone is really cool, the display is just gorgeous. The battery life is really bad, however."

"Oh, of course - I have a lot of time. Just keep on using my money for paying those really fast and friendly authorities."

Obviously, this hardens the task of sentiment analysis. Moreover,it is hard to teach a machine patterns like sarcasm and to identify the intended meaning behind words. [2, 4]

Another crucial point are dependencies of sentiments: topic, domain, and temporal dependency. Those mainly focus on the problem, that sentiments can have a different meaning, depending on the underlying topic or domain. E.g. a word

[1] `http://twitter.com`

such as "unpredictable" may have a positive meaning if it used for movie review, but could have a negative sentiment if it used to describe the behavior of a car. Temporal dependency deals with training a classifier with data from a certain time-period and use this classifier for data of another time-period. This may may have an influence on the accuracy of classification as well. [5]

2.2 Twitter-based Data

Twitter is a form of microblog, where users can post small text immediately and the so called tweets are contain real-time reactions to certain events. The social network platform is categorized as microblog as every tweet is limited to 140 signs. This results in people using e.g. abbreviations, emoticons, (intentional) spelling mistakes in order to fit and express their opinion accurate. Moreover, Twitter uses some special characters like the @ which indicates that the post directed to another user. In addition hashtags are used to refer to special topic. All these special characteristics play a huge role when analyzing tweets. [1, 5]

3. TWITTER API

Tweets has become to a popular resource for sentiment analysis, as Twitter provides an API to retrieve tweets and therefore makes the collection of data easy.

4. PREPROCESSING TWEETS

In this section we will focus on how to conduct sentiment analysis for tweets and suggested approaches.

5. CLASSIFYING TWEETS

5.1 Corpus

The corpus is the starting point of each sentiment analyses. It contains the data the will be used to train a classifier.

5.2 Training

5.3 Evaluation

6. RELATED WORK

7. CONCLUSION

8. REFERENCES

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[2] B. Liu. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*, 2010.

[3] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[4] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[5] J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.