State of the Art in *Twitter Sentiment Analysis*

Christian B. Hotz-Behofsits
0929002
christian.hotz-behofsits@tuwien.ac.at

Thomas Schmidleithner
1025525
tschmidleithner@auto.tuwien.ac.at

Dominik Pichler
1026045
dominik.pichler@aon.at

Matthias Reisinger
1025631
matthias.reisinger@web0.at

Florian Taus
0627918
florian.taus@hotmail.com

ABSTRACT

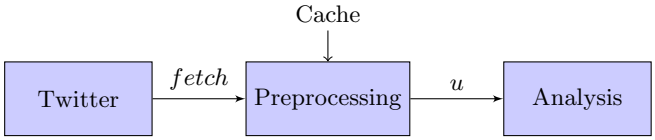
The extensive growth of user-generated content in social networks and the common usage of emoticons and hashtags has introduced new possibilities to classify these information. In this paper, an overview of the state-of-the-art regarding sentiment analysis of Twitter messages is provided. The usefulness of existing lexical ressources and special expressions like smileys or hashtags is evaluated and an general introduction to sentiment analysis is given as part of the introduction.

1. INTRODUCTION

Within the last years Twitter and similar social media platforms grew considerably in terms of user numbers and mainstream fame. Twitter has about 284 Mio. active users, generating approximately 500 Mio. posts[1] on a single day[2]. These impressive numbers show the amout of information generated by the crowd. Furthermore Twitter is able to reach a vast number of potential customers especially for consumer businesses. As a result the named services gain more and more acceptance as opinion platforms and powerful tools for both, opinion leader as well as pollster. This trend influenced various marketing and sales strategies and created a new market for companies specialized on sentiment analysis of tweets, like tweetfeel[3], Social Mention[4] and Twitratr[5].

The aim of sentiment analysis is to determine and try to measure positive and negative feelings, emotions and opinions written in a text. English as language allows to express the same intent in different ways. The main challenge therefore consists in abstracting the intention of the writer from the grammatical and language specific rules.

In general, sentiment analysis can be splitted into two steps: a preprocessing and a constitutive sentiment analysis phase. In Twitter sentiment analysis there is another step right before preprocessing: fetching the data from the application programming interface (API). This task is not trivial, because the official API is limited regarding the datasets fetched within certain intervals. This leads to the requirement of a caching mechanism to avoid fetching the same tweets multiple times and allow the usage of more datasets.



While sentiment analysis of conventional resources like news papers or articles is quite well investigated the special research area of sentiment analysis in terms of social network posts is relatively unexplored. Especially the possibility of priorizing keywords[6] and the common usage of emoticons lead to advanced possibilities to categorize feelings of posts.

2. RELATED WORK

Sentiment analysis a growing and well explored part of Natural Language Processing (NLP). There are several papers from Pang and Lee regarding sentiment analysis in general and related topics like the effects of various machine learning approaches [10][9]. Especially the last topic is a well studied field [8]. Some related research areas are document level classification, sentence-level classification and machine learning. Due the limited number of characters within a single twitter post, this topic is quite similar to sentence-level sentiment analysis. There are several recent papers regarding sentiment analysis in the context of twitter posts. Agarwal et al. published within "Sentiment analysis of Twitter data" general information and techniques which cover this area [1]. In "Twitter Sentiment Analysis : The Good the Bad and the OMG !" the utility of linguistic features for detecting the sentiment of tweets are investigated. Saif, He, and Alani introduce a novel approach of adding semantics as additional features into the training set. For each extracted entity (e.g. Galaxy S) from Twitter Messages, they add a semantic concept (e.g. Samsung product) as an additional feature, and measure the correlation of the representative concept with negative/positive sentiment [11]. In *Twitter Sentiment Analysis*, an algorithm is presented, which accurately classifies tweets as positive or negative in respect to

[1] posts are also known as tweets
[2] `https://about.twitter.com/company`, Effective 12.11.2014
[3] `www.tweetfeel.com`
[4] `www.socialmention.com`
[5] `www.twitratr.com`
[6] known as hashtags

a query term[4].

3. DATA

Twitter is a social network and microblogging service that allows their users to compose and read so called *tweets*. The user registration is not limited to professional writers, authors or similar expert groups and there are no quality safeguards or writing guidelines.

Each tweet is a text message and can contain up to 140 characters including links, pictures[7], and special tagged words. These special tagged words start with a well defined prefix and add semantic information to the prefixed string. Table 1 shows an overview of common prefixes supported by Twitter and their purpose. A hash (#) labels a word, also known as hashtag. It is used to mark important keywords within tweets. There is also a at-sign-prefix @, which allows to mention and link other Twitter users in a post. Tweet composers can insert URLs as part of their messages. Regarding the hard length limitation of a tweet, url-shortener are often used to insert longer internet addresses.

prefix	name of character	meaning
#	hashtag	keyword, index
@	at-sign	twitter username

Table 1: Common Twitter word prefixes

The combination of the 140 character restriction and the missing writing guidelines leads to another problem: Slang, emoticons and punctuation that are uncommon in longer text messages are heavily used.[2] In terms of sentiment analysis, tweet processing is more comparable to sentence level analysis as paragraph analysis.

4. PREPROCESSING

The data preprocessing step is an essential part of sentiment analysis. Its goal is to prepare data for the sentiment analysis and remove noisy, unnecessary, inconsistent or incomplete parts.[5] In a first step a tokenizer splits the tweet into conjugate parts, called words. The resulting list of words is filtered regarding the word class, the word itself and the intention of the sentiment analysis. In case of sentiment analysis of twitter posts it is common to remove the parts described in the following paragraphs.[5]

URLs. Uniform Resource Locators (URLs) refer to another site, which in fact can contain very important information, but the internet address on its own is not of interest, because in most cases it contains only auto-generated information, which can not be influenced by the poster.

Unnecesarry words. It is not an obvious task to classify if a word is unnecessary or not. But some word classes, like pronouns, articles or interrogatives do not contain any relevant information and can be skipped. (TODO:Referenz)

[7]pictures are replaced by a url targeting a public available image resource

Retweets. Retweets reflect the position of another person and are not always created to express the same opinion (e.g. statement of sarcasm).

Repeated characters. Tweets contain often deformed words containing repeated characters like 'heeeey' or 'whuhuuu'. In this cases it is intended to reconstruct the correct english term and only if it is not possible or too sophisticated, the word is removed by the preprocessor.

4.1 Tokenizer

There are already several open source tokenizer implementations available. Some popular programming libraries for the english language include the Stanford NLP PTBTokenizer [8] and the Apache OpenNLP Tokenizer[9]. These examples are quite versatile and can even be used for different tasks and languages. To overcome the special needs of sentiment analysis in terms of Twitter posts, highly specialized Tokenizer were created (e.g. Carnegie Mellon Twokenizer[10]). Some tokens like emoticons are difficult to handle. Although they are a good and popular resource for sentiment analysis. [6] It is a sophisticated task to recognize them correctly and some Tokenizers like the Stanford NLP PTBTokenizer even skips them per default. The Carnegie Mellon's Twokenizer recognizes the related parts and forwards them to the Tagger.

4.2 Tagger

Part-of-Speech (POS) tagging is a simple form of syntactic analysis and therefor very useful for Natural Language Processing. The performance of POS taggers degrade on out-of-domain data and tagging tweets poses additional challenges, e.g. the twitter specific slang, emoticons and the lack of conventianal orthography. [3] There is already a wide range of freely downloadable POS taggers [11], e.g. the Standford POS tagger [12]. Because of the peculiarities of the language used in tweets, just a few of them are specialized on tagging twitter data, like the TweetNLP tagger [13]. In comparison with the Stanford POS tagger the TweetNLP tagger reduces errors by 25%. [11]

5. CLASSIFICATION

Sentiment analysis of twitter data, or more general, the idea of extracting sentiment and opinions from pieces of text is based on the more general principle of *classification* [10]. The broad aim of classification is to assign given textual units to a set of classes or categories or the apply some kind of regression or ranking. Sentiment analysis is a specialization of this approach which aims at assigning sentiment values to documents. One application might involve to classify an opinionated text by assigning one of two opposing sentiment polarities, i.e. classifying it as either *positive* or *negative*. This classification task is also referred to as *binary classification* or *sentiment polarity classification* [10]. But

[8]http://nlp.stanford.edu/software/tokenizer.shtml
[9]http://opennlp.apache.org/documentation/manual/opennlp.html
[10]http://www.ark.cs.cmu.edu/TweetNLP/
[11]http://www-nlp.stanford.edu/links/statnlp.html
[12]http://nlp.stanford.edu/software/tagger.shtml
[13]http://www.ark.cs.cmu.edu/TweetNLP/

in general, the input to the sentiment classification process is not strictly opinionated which makes this task challenging. Therefore this kind of binary classification might not always be applicable. Different approaches that allow for a more fine grained classification might be appropriate, for example, based on a multi-point scale that allows for more than just two sentiment classes.

The characteristics of twitter messages introduce further challenges. Due to the structure of these messages, sentiment analysis of tweets is different from analyzing conventional texts, such as review documents, in various ways. The length of at most 140 characters and the rather informal spelling style pose problems that have to be considered carefully when analyzing the data. When preprocessing the texts these aspects already need to be handled so that the actual classification process is supplied with information that is considered useful for the analysis phase. After this step it is necessary to collect the relevant information as so called *features* and organize them into a *feature vector*.

5.1 Features

Since twitter messages are composed of an arbitrary number of words, classification makes it necessary to extract exactly those words which preferably describes the characteristics of such tweets. The extraction of those words by the use of various word selecting techniques is called a feature vector. These word selecting techniques are a fundamental process for providing informations to the classification methods, thus an extensive body of different word selecting algorithms exist. An overview is provided by Pang et al. [9] which can be summarized as follows:
Term Presence is a word selecting technique where the feature vector consists of a binary value, 0 for "term is not present" and 1 for "term is present".
A similar approach is to count the number of occurred words in a message and make use of the numeric value as a feature, which is called *Term Frequency*.
Term-based Features Beyond Term Unigrams describes the encoding of position informations within a textual unit (e.g., at the end of the document) into the feature vectors. This technique is based on the idea that the position of textual units can have influence on the importance of these.
The idea of *Part-of-Speech* (POS) informations is to make use of labeled words within the tweet by diversifying them into categories of words. As soon as the words are labeled, they can be filtered (f.e. by adjectives and verbs only which can be a strong indicator for sentiment) and the result can be applied to the feature vectors.

5.2 Methods

For the actual classification process extent literature distinguishes two general types of techniques, namely *machine learning* and *semantic orientation* [12]. The machine learning approach is also referred to as *supervised learning*, since corresponding techniques generally require supervised training phases. Accordingly, semantic orientation is also called *unsupervised learning* because such a training phase is not necessary for the process to work. The most prominent machine learning methods comprise *support vector machines* and the *Naïve Bayes* classifier.

Naïve Bayes

This technique is based on a simple stochastic model that for a given document d tries to find the most likely class $c^* = \text{argmax}_c P(c|d)$ where c stands for one out of a number of possible classes and $P(c|d)$ denotes the conditional probability that document d can be assigned to c [10]. For example, in sentiment analysis c^* would denote the sentiment value that could be assigned to document d. c^* is computed by applying Bayes' rule,

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}$$

for each possible sentiment value c. $P(d|c)$ has to be estimated and therefore it is presumed that there exists a predefined set of features $\{f_1, \ldots, f_m\}$ which are assumed to be conditionally independent. Based on this estimation the actual rule that can be used for classification is derived from Bayes' rule:

$$P_{\text{NB}}(c|d) := \frac{P(c)(\prod_{i=1}^{m} P(f_i|c)^{n_i(d)})}{P(d)}.$$

Here, $n_i(d)$ designates the number of occurrences of feature f_i in document d.

Support Vector Machines

In contrast to the above method, support vector machines (SVMs) do not use a probabilistic model. Instead the aim of this technique is to find a so called *hyperplane*, which is represented by a vector $\vec{\omega}$. Let $\vec{d} := (n_1(d), n_2(d), \ldots, n_m(d))$ be the *document vector* of document d where, like above, $n_i(d)$ represents the number of occurrences of feature f_i in d. Then, in the case of a binary classification problem, $\vec{\omega}$ separates the document vectors in one class from the vectors in the other class such that this separation (the so called *margin*) is as large as possible. To find $\vec{\omega}$, a dual optimization problem has to be solved, to find for each document d_j a value α_j such that the following holds:

$$\vec{\omega} := \sum_j \alpha_j c_j \vec{d_j}, \quad \alpha_j \geq 0,$$

where $c_j \in \{1, -1\}$ (in sentiment classification 1 would denote "positive" and correspondingly -1 stands for "negative" sentiment). The actual classification of document d is done, by determining on which side of $\vec{\omega}$ its document vector $\vec{d}$ falls on.

Further techniques

Beside Naïve Bayes and Support Vector Machines various other machine learning techniques such as Character based N-gram model, Maximum entropy, Random Forest and J48 exist and are just mentioned here for the sake of completeness, a comprehensive view of other machine learning techniques is beyond the scope of this survey.

6. CONCLUSIONS

Due to the emerging number of users and Tweets, Twitter sentiment analysis is a growing field. Sentiment analysis itself is a well explored part of NLP, but the analysis of Twitter posts is not as widely investigated as for conventional resources like newspapers or articles.

However, the analysis of tweets is more like the sentence-level than the paragraph-level analysis due to their short length, which also leads to other problems as the use of slang, emoticons and other uncommon sentence parts. To prepare the data, URLs, unnecessary words, retweets and repeated characters are removed in the preprocessing phase. The use of the right tagger, which ought to be a tagger specialized for Tweets, is important because of the big differences, e.g. TweetNLP has 25 percent less errors than the Stanford POS tagger.

The relevant information ("features") is then determined and classified with the support vector machine or the Naïve Bayes method. The Naïve Bayes method classifies documents based on the Bayes' rule while support vector machines are trained to create a hyperplane that separates the two classes ("positive" and "negative" in sentiment classification), so that Tweets can be classified as above or below this hyperplane.

We saw that Twitter sentiment analysis is an important research field but needs some more work to improve the quality and accuracy of both preprocessing and classifying the content of a tweet.

APPENDIX

References

[1] Apoorv Agarwal et al. "Sentiment analysis of Twitter data". In: *Association for Computational Linguistics* (2011), pp. 30–38.

[2] Alexander Davies and Zoubin Ghahramani. "Language-independent Bayesian sentiment mining of Twitter". In: (2011).

[3] Kevin Gimpel et al. "Part-of-speech tagging for Twitter: annotation, features, and experiments". In: *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*. Vol. 2. 2011, pp. 42–47.

[4] Alec Go, Lei Huang, and Richa Bhayani. *Twitter Sentiment Analysis*. Tech. rep. 2009, p. 17.

[5] Govardhan Hemalatha Saradhi-Varma. "Preprocessing the Informal Text for efficient Sentiment Analysis". In: *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)* 1 (2012), pp. 58–61.

[6] Alexander Hogenboom et al. "Exploiting emoticons in sentiment analysis". In: *Proceedings of the 28th Annual ACM Symposium on Applied Computing*. ACM. 2013, pp. 703–710.

[7] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. "Twitter Sentiment Analysis : The Good the Bad and the OMG !" In: *Artificial Intelligence* (2011), pp. 538–541.

[8] Christopher D Manning and Hinrich Schütze. *Foundations of Natural Language Processing*. 2000, p. 678.

[9] Bo Pang and Lillian Lee. *Opinion Mining and Sentiment Analysis*. 2008. arXiv: `0112017 [cs]`.

[10] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques". In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*. Vol. 10. Association for Computational Linguistics, 2002, pp. 79–86.

[11] Hassan Saif, Yulan He, and Harith Alani. "Semantic sentiment analysis of twitter". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7649 LNCS. PART 1. 2012, pp. 508–524.

[12] Qiang Ye, Ziqiong Zhang, and Rob Law. "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches". In: *Expert Systems with Applications* 36.3, Part 2 (2009), pp. 6527 –6535.