State of the Art in *Twitter Sentiment Analysis*

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

Thomas Schmidtleitner
1025525
e1025525@student.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

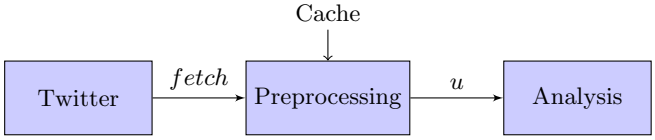
The extensive growth of user-generated content in social networks and the common usage of emoticons and hashtags has introduced new possibilities to classify these information. In this paper, an overview of the state-of-the-art regarding sentiment analysis of Twitter messages is provided. The usefulness 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 pollister. This trend influenced various marketing and sales strategies and created a new market for companies specialiced 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 grammatic and language specific rules.

In general sentiment analysis can be splitted into 2 steps: a preprocessing and a constitutive sentiment analysis phase. In Twitter sentiment analysis there is another step right bevor 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 conventinal ressources 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. Excpecially 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 Langugage 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[8][7]. Especially the last topic is a well studied field [6]. 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 sentance-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 de-

[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

tecting 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[9]. In *Twitter Sentiment Analysis* a algorithm is presented, which accurately classifies tweets as positive or negative, in respect to a query term[2].

3. DATA

Twitter is a social network and a microblogging service that allows their users to compose and read so called *tweets*. Each tweet is a text message and can contain up to 140 characters including links, pictures[7], and special tagged words. These special words start with a well defined prefix and add semantic information to the prefixed word. Table 1 shows an overview of common prefixes supported by Twitter and their purpose. The hashtag # tags a special word, also known as hashtag. It is used to mark important keywords within the tweet. There is also a at-sign-prefix @, which allows to mention other Twitter users in a post. It's also possible to insert urls as part of the message. Regarding the hard message length limitation of a tweet, url-shortener are often used to insert longer urls.

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

Table 1: Common Twitter word prefixes

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.[3] [4] 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. [3]

URLs. Uniform Resource Locators (URLs) refer to another site, which in fact can contain very important information, but the url itself is not of interest, because in the most cases it contains autogenerated information, which can not be influenced by the poster.

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

Retweets. Retweets reflect the postion of another person and are not always created to express the same position (it can also be a statement of sarcasm).

Repeated characters. Everybody can post on Twitter, it is not limited to professional writers, authors or similar expert groups. Therefore the 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 to 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 be even used for different tasks and languages. The restirction of 140 Characters leads to a heavy usage of slang, emoticons and punctuation that are uncommon in longer text messages.[**davies2011**] This circumstances give rise to specialized Tokenizer for tweets. One example is the specialized Carnegie Mellon Twokenizer[10].

It is a sophisticated task to recognize smileys. Some tokenizers.

Some words are difficult to handle. One example are emoticons, because they are hard to recognize correctly. Although they are a good and popular ressource for sentiment analysis. [**emoticons**]

4.2 Tagger

5. METHODS

6. CONCLUSIONS

APPENDIX

A. HEADINGS IN APPENDICES

References

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

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

[3] 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.

[4] I Hemalatha, GP Saradhi Varma, and A Govardhan. "Preprocessing the Informal Text for efficient Sentiment Analysis". In: ().

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

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

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

[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/

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

[8] 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.

[9] 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.