Poster: Sentiment Analysis of Twitter Data:

Towards Filtering, Analyzing and Interpreting Social Network Data

Lisa Branz
Department of Computer Science
Technische Hochschule Nuremberg Georg Simon Ohm
Nuremberg, Germany
branzli56977@th-nuernberg.de

Patricia Brockmann
Department of Computer Science
Technische Hochschule Nuremberg Georg Simon Ohm
Nuremberg, Germany
patricia.brockmann@th-nuernberg.de

ABSTRACT

Social networks provide a rich data source for researchers that can be accessed in a comparatively effortless way. As data and text mining methods such as Sentiment Analysis are becoming increasingly refined, the wealth of social network data opens up entirely new possibilities for exploring specific in-depth research questions. In this paper an approach towards the retrieval, analysis and interpretation of social network data for research purposes is developed. The data is filtered according to relevant criteria and analyzed using Sentiment Analysis tools tailored specifically to the data source. The approach is verified by applying it to two example research questions, confirming past findings on cultural and gender differences in sentiment expression.

CCS CONCEPTS

• **Computing methodologies → Information extraction**;

KEYWORDS

Big Data, Data Mining, Sentiment Analysis WEKA, Twitter, Social Networks, Social Media, Global Software Engineering

ACM Reference Format:
Lisa Branz and Patricia Brockmann. 2018. Poster: Sentiment Analysis of Twitter Data:: Towards Filtering, Analyzing and Interpreting Social Network Data. In *DEBS '18: The 12th ACM International Conference on Distributed and Event-based Systems, June 25–29, 2018, Hamilton, New Zealand*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3210284.3219769

1 INTRODUCTION

The widespread growth of social networks throughout the past decade has opened up entirely new possibilities for researchers when it comes to collecting large amounts of data. With millions of conversations taking place on social networks every day they offer a rich data source that can be accessed in a comparatively effortless way. The microblogging service Twitter [3] allows developers to connect to its Streaming API to receive a real-time data stream containing Twitter posts ("tweets") and user information. The brevity of the posts as well as their mostly text-based nature facilitate the analysis using data mining methods and have, therefore, made Twitter a popular data source for scientific research. As data mining and text mining techniques become more advanced, information can be retrieved on an increasingly fine-grained level. An area that has received considerable attention throughout the past years is Sentiment Analysis. The method has been widely applied to capture individuals' sentiment towards products or to assess the overall sentiment expressed in a piece of text. Besides the more general classification of comments or reviews as positive, neutral or negative, it also allows researchers to identify the type and intensity of more distinct emotions, such as fear, joy or surprise, in written text. However, being able to detect emotional content in social network data does not necessarily imply that useful knowledge can be derived from it. Despite its benefits in regards to wealth and accessibility, social network data is naturally unstructured and noisy. It is therefore not a trivial task to filter and collect large amounts of data relevant to a specific research question and to choose appropriate tools for further analysis in order to obtain meaningful results. In this paper an approach for acquiring, analyzing and interpreting suitable data from social networks using Sentiment Analysis is developed. Social network data from a research project on gender and cultural diversity in global software engineering is then used to verify the approach by applying it to two example research questions.

DEBS '18, June 25–29, 2018, Hamilton, New Zealand

ACM ISBN 978-1-4503-5782-1/18/06.
https://doi.org/10.1145/3210284.3219769

2 RELATED WORK

The collection and Sentiment Analysis of social network data has already been addressed by multiple papers (e.g., [8, 9, 11–15, 18]). In this section, we will introduce the tools that were used for this approach as well as previous findings regarding the specific research questions in this paper.

2.1 Retrieval of Tweets

There are a number of different ways to access the Twitter Streaming API and to consume a stream of real-time tweets. A list of libraries for different programming languages is provided by Twitter [4]. For the present work the Twitter4J Java library was used to connect to the Streaming API and collect real-time tweet data [5].

2.2 Sentiment Analysis

For the specific requirements in this paper a tool specifically designed for detecting sentiment in Twitter data was chosen. The area has received considerable attention in recent years with an extensive amount of work conducted towards developing tools for the reliable classification of sentiment in Twitter posts [8, 9, 12, 13, 15]. For the present research WEKA [10] was used to conduct Sentiment Analysis on the collected tweet data. The open source software

published and maintained by the Machine Learning Group at the University of Waikato offers a large variety of machine learning algorithms for data mining tasks. The AffectiveTweets Package for the WEKA workbench developed by Bravo-Marquez and his team [18] was used to determine the types of emotions and their intensity in tweet content. The collection of unsupervised filters was developed to extract features for sentiment classification from a tweet, taking into account aspects such as emoticons, the number of words and expressions indicating a certain emotion, and hashtags used to intensify the expressed emotion. The tool includes manually created sentiment lexicons as well as tweet-specific, automatically generated lexicons [14].

2.3 Previous Findings

In recent years an increasing number of studies have used Sentiment Analysis in order to answer specific research questions. These findings suggest that Sentiment Analysis is indeed an effective method for answering specific research questions using data gathered from computer-mediated conversation. Yang et al. analyzed e-mail and instant messaging conversations among employees from seven different countries and found cultural differences regarding the expressed sentiment [11]. These findings of Yang et al. provided the base for one of the exemplary research questions in this paper. In a field study Faulkner [16] found that female engineers often feel left out in male-dominated work places. One of the reasons might be the conversation topics in these workplaces, as they often appear to be dominated by stereotypically male interests, such as sports. This paper aims at confirming Faulkner's results by applying Sentiment Analysis to social network data.

3 OBJECTIVES

The aim of this paper is to present an approach towards acquiring, filtering, analyzing and interpreting social network data from Twitter with regard to specific sentiment-related research questions. Focus is laid on acquiring data from a specific user group while excluding irrelevant data from the data pool. Unavailable information that is crucial for the research questions at hand is inferred from the collected data. The approach is verified by exploring two example research questions from the software engineering field. Based on the findings of Yang et al. [11] and Faulkner [16] the following two hypotheses are proposed:

Hypothesis 1: The amount of positive sentiment in tweets about sports differs significantly between male and female software engineers, with males showing more positive sentiment towards sports-related topics than females.

Hypothesis 2: The amount of sentiment expressed in tweets differs significantly between software engineers from collectivist cultures and software engineers from individualist cultures, with users from collectivist cultures expressing less sentiment.

In the following section, we will present an approach towards overcoming four main challenges with regards to deriving meaningful information from large amounts of unstructured social network data.

3.1 Filtering Relevant Posts

While many studies in the past have focused on collecting large amounts of data, only few studies have focused on collecting social network data from very specific user groups such as software engineers. As the research questions in this paper specifically focus on the software engineering community, the aim was to filter the real-time Twitter stream accordingly and to exclude large amounts of irrelevant data.

3.2 Inferring Information from Collected Data

Not all information relevant to a specific research question can be collected directly from the social network. Some information crucial to the research questions at hand, such as a user's gender or cultural background, had to be inferred from the data available.

3.3 Sentiment Analysis of Tweet Text

Emotion types and intensities had to be extracted from every tweet text in the sample. In order to allow for comparison, a fine-grained Sentiment Analysis was conducted, identifying not only the specific types of emotions present in a tweet but also their intensity as a quantifiable value.

3.4 Statistical Analysis of Sentiment Data

The amount of general sentiment as well as specific emotions expressed in tweets were compared between user groups using statistical methods in order to confirm the differences postulated in the hypotheses.

4 METHOD

4.1 Data Sample

We collected Twitter data via the Twitter Streaming API. The Twitter4J Java library [5] was used to access the API and retrieve a real-time stream of tweets and their respective author's name, profile description and location. For the data sample two separate sets of tweets were collected. The first set contained tweets related to software engineering topics. The second set contained tweets posted by users who work in software engineering, which could revolve around any topic. In order to retrieve tweets about software engineering topics, a list of keywords related to the topic was developed using the Related Words tool [2]. Only tweets containing at least one of the keywords were collected. A similar approach was adopted for tweets posted by users who work in software engineering. The code was modified to display any tweets posted by a user whose profile description contained a software engineering-related keyword also found on a keyword list.

A bounding box approach was adopted using latitude and longitude values in order to only collect tweets sent from specific geographic regions. The regions were selected based on Hofstede's cultural dimensions [17], in order to allow for comparison between collectivist and individualist cultures. Based on their score on Hofstede's Individualism dimension, Canada, the US, Ireland, the UK and Australia were classified as individualist cultures, whereas India and South East Asia were classified as collectivist cultures. 53.7% were collected from individualist cultures, whereas 46.3% of tweets were collected from collectivist cultures.

As a Twitter user's gender is not accessible through the Twitter API, this information had to be determined using the information available. Using Gender API [1], a user's gender was determined from their first name. All predictions with an accuracy of .80 or higher were accepted. The overall gender ratio was then checked for plausibility using current statistics from industry [6, 7]. 85.9% of tweet authors were classified as male whereas 14.1% were classified as female. In total, 1028 tweets were collected.

The tweets posted by users in software engineering jobs were then classified for tweet topic. For the hypotheses proposed, a precise and faceted classification of tweet topics was not necessary. Hence, in order to test Hypotheses 1, it was considered sufficient to classify a tweet as sports-related or non-sports-related. The tweet set was hand-annotated with the tweet topic, assigning a tweet either to the "Sports" category or to the "Other" category. Tweets that were assigned to the "Sports" category included posts about watching sporting events, discussing teams or sports personalities or mentioning one's own exercise-related activities. All tweets that did not match one of these criteria were classified as "Other".

4.2 Sentiment Analyis Using WEKA

Sentiment Analysis was performed on the collected tweet data using WEKA [10]. The AffectiveTweets Package [18] for the WEKA workbench was used to determine the types of emotions and their intensity in tweet content. Applying the TweetToLexiconFeatureVector filter to the tweet content in WEKA added 43 feature attributes to the data file, which contained numeric scores that indicate the strength of detected positive and negative sentiment in general, as well as specific emotions: anger, anticipation, disgust, fear, joy, sadness, surprise and trust. The sentiment scores are based on several sentiment lexicons.

4.3 Statistical Analysis

In order to test Hypothesis 1 all tweet authors whose gender was determined as "Unknown" were excluded from the analysis. Using an independent sample t-test, the means of all sentiment scores of males and females were compared for all topics to create a baseline. The procedure was then repeated for sports-related tweets only. In order to test Hypothesis 2, the means of all sentiment and emotion scores of users from individualist and collectivist cultures were compared using an independent sample t-test.

5 RESULTS

5.1 Gender-Specific Preference for Conversation Topics

In order to test Hypothesis 1, the sentiment scores of male and female users were compared for all tweets in the tweet set (n=893). A t-test revealed significantly higher scores for females across seven out of 42 sentiment scores, all of which belonged to the positive sentiment range. Males scored significantly higher on NRCHashSentnegScore (p=.040). None of the other scores differed significantly between males and females. It can, therefore, be concluded that the women in the sample tend to express more overall positive sentiment in their tweets than the men. Another t-test was performed only taking into account the 108 tweets in the "Sports" category. The significant differences between males and females that were found in the general sample disappeared when looking only at sports-related tweets. Males showed significantly higher scores for all feature attributes in Table 1. No sentiment score was significantly higher in females than in males for sports-related tweets. Hypothesis 1 could therefore be confirmed. Additionally, men were also found to express more negative sentiment towards sports than women.

Table 1: Sentiment Scores of Male and Female Users for Sports-Related Tweets

Sentiment Category	Mean value		p (1-tailed)
	Male	Female	
S140negScore	-1.7056	-.6437	.0355
NRC10job	.33	.00	<.0001
NRC10surprise	.16	.00	<.0001
NRC10Expandedangry	.6485	.3658	.0495
NRC10Expandedsad	.2700	.1024	.0025
NRC10HashEmoant	.9965	.4735	.048

5.2 Cultural Differences in Expressed Sentiment

Although not all results were significant, they showed a very clear pattern. Users from individualist cultures scored higher on 40 out of 42 feature attributes, with the two exceptions being non-significant. In the present analysis, software engineers from collectivist cultures consistently showed less sentiment than users from individualist cultures. It is also noteworthy that tweet authors from collectivist cultures used significantly fewer negations in their tweets (p=.018). Hypothesis 2 could be confirmed. The results suggest that software engineers from collectivist cultures indeed express less sentiment in their posts than users from individualist cultures. The findings of Yang et al. [11] could be replicated following the approach proposed in this paper.

6 DISCUSSION

In summary, the results confirm both hypotheses proposed. The findings of Faulkner and Yang et al. could be confirmed and extended using Sentiment Analysis on Twitter data. It can, thus, be concluded that the approach presented is indeed suitable for acquiring, filtering, analyzing and interpreting social network data in order to explore specific research questions regarding sentiment in social network communication. Further, the approach does not only allow us to retrieve and analyze large amounts of Twitter data, it also presents a way to extract data of a very specific user group from unstructured and noisy social network data with comparatively little effort. It is, therefore, suitable for research questions focusing on a broad as well as a narrowly circumscribed user group.

6.1 Method

Data was collected from Twitter, the largest and most popular microblogging platform to date. This approach was adopted due to the availability of a large number of posts from users all around

the globe as well as already established ways of retrieving data. However, there are some downsides to this approach. The data was selected by filtering tweets and user descriptions that contained software engineering-related keywords. The keywords were selected carefully to exclude as many irrelevant tweets as possible. However, this approach still carries the risk of retrieving a certain amount of data not relevant to the research question. A way to avoid this problem could be to collect data from corporate social networks. While this approach considered, it was eventually abandoned due to strict data privacy regulations, the smaller amount of data available as well as the possibility of biased data.

6.2 Data Sample

The adequacy of the approach presented in this paper is further supported by the fact that a large majority of the collected tweets were posted by male users while only a comparatively small portion was posted by female users. While this might not be ideal for statistical analysis, it reflects very accurately the gender distribution in software engineering. Taking into account that the presumed gender ratio on Twitter is generally very balanced [19], the imbalanced gender ratio in the sample supports the assumption that the overall selection of tweet data was valid.

A bounding box approach was adopted in order to collect tweets sent from locations within the specified boundaries. Especially when it comes to the cultural background of a Twitter user it is important to note that not everyone who is sending a tweet from a specific location is necessarily living in the area. It is almost unavoidable to collect a certain number of tweets from users who do not live in the place they tweeted from. However, in a vast majority of cases the bounding box location of a user matched the location stated in their profile. Despite its flaws, the bounding box approach still appears to be the most reliable way to retrieve valid information on the location and, thus, the cultural background of a tweet's author.

6.3 Flexibility

A similar approach using different keyword lists for both tweet content and profile descriptions could be adopted for gathering data from other user groups in order to explore a variety of different research questions. The approach presented in this paper is, however, specifically tailored to the analysis of Twitter data. Sentiment Analysis tools developed for other data sources would be required in order to adapt this approach to other social networks.

7 CONCLUSION AND FUTURE WORK

An approach towards acquiring, filtering, analyzing and interpreting Twitter data using Sentiment Analysis was developed in this paper. By confirming previous findings regarding two research questions from the field of software engineering, the validity of the approach could be demonstrated.

More than 1000 tweets were collected for this research, in order to develop a working and verified approach. Data mining methods, however, yield the possibility of confirming the present findings based on an even larger amount of social network data. In order to make use of the full potential of data mining methods, the collection of larger data samples should be pursued in the future. The amount of data could be increased by collecting tweets over a longer period of time or through the additional use of the Twitter Search API.

An automated classification of tweet topics specifically tailored to the research question at hand would be desirable for future studies with larger data samples.

Using data from corporate social networks could help overcome some of the challenges regarding the collection of valid data.

ACKNOWLEDGMENTS

This work was supported by a grant from Technische Hochschule Nuremberg as part of the research project "Diversenta".

REFERENCES

[1] 2018. Gender API. Retrieved February 10, 2018 from https://gender-api.com/
[2] 2018. Related Words - Find Words Related to Another Word. Retrieved January 11, 2018 from http://relatedwords.org/
[3] 2018. Twitter. ItâĂŹs whatâĂŹs happening. Retrieved January 11, 2018 from https://twitter.com/
[4] 2018. Twitter libraries âĂŤ Twitter Developers. Retrieved March 4, 2018 from https://developer.twitter.com/en/docs/developer-utilities/twitter-libraries
[5] 2018. Twitter4J - A Java library for the Twitter API. Retrieved February 11, 2018 from http://twitter4j.org/en/index.html
[6] C. Ashcraft. 2016. WOMEN IN TECH: THE FACTS. Retrieved February 6, 2018 from www.ncwit.org/thefacts
[7] J. McGrath Cohoon. 2003. Must there be so few? Including women in CS. In *Proceedings of the 25th International Conference on Software Engineering*.
[8] E. MartÃŋnez-CÃąmara et al. 2014. Sentiment analysis in Twitter. *Natural Language Engineering* 20, 1 (2014), 1–28.
[9] F. Bravo-Marquez et al. 2014. Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems* 69 (2014), 86–99.
[10] I. H. Witten et al. 1999. *Weka: Practical machine learning tools and techniques with Java implementations.* Working Paper 99/11. University of Waikato, Department of Computer Science, Hamilton, NZ.
[11] J. Yang et al. 2011. Collaborating Globally: Culture and Organizational Computer-Mediated Communication. In *Proceedings of the International Conference on Information Systems.*
[12] Kiritchenko et al. 2014. Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research* 50 (2014), 723–762.
[13] M. Thelwall et al. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology* 63, 1 (2012), 163–173.
[14] S. Mohammad et al. 2013. NRC-Canada Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises.*
[15] S. Mohammad et al. 2017. Stance and Sentiment in Tweets. *ACM Transactions on Internet Technology* 17, 3 (2017), 1–23.
[16] W. Faulkner. 2009. Doing gender in engineering workplace cultures I. Observations from the field. *Engineering Studies* 1, 1 (2009), 3–18.
[17] Geert Hofstede. 1984. *CultureâĂŹs Consequences: International Differences in Work-Related Values.* SAGE.
[18] S. Mohammad and F. Bravo-Marquez. 2017. Emotion Intensities in Tweets. In *Proceedings of the Joint Conference on Lexical and Computational Semantics.*
[19] Aaron Smith. 2018. Social Media Use in 2018. Retrieved February 25, 2018 from http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/