Sentiment Analysis of Short Texts on Twitter Data

Abstract

In this paper, we study the sentiment analysis of short text data from Twitter by extracting the information from tweets and tweet replies. Our method concerns contents and sentiments of tweets and tweet replies. We have collected more than 21 gigabytes of Twitter data to design and analyze the sentiments of short texts more precisely when compared with conventional methods of sentiment analysis. We have experimented with different datasets to implement our model. Finally, we have developed a dataset with tweets and tweet replies along with different hand-crafted features. In the next step, different machine learning algorithms are used to train the model. Naive Bayes, Decision Tree, SVM, Random Forest, and NLTK are used to build the model. SVM outperformed the other algorithms. The new methodology implemented to handle twitter data performed well and showed a significant increase in accuracy. Therefore, our approach analyzes the sentiments of short and ambiguous texts more accurately by iteratively appending tweets and tweet replies of selected topics in Twitter.

Keywords

Twitter, Sentiment, Replies, Short Texts, SVM, NLTK, Random Forest

1. Introduction

Microblogging today has become a popular communication tool among internet users. Millions of users share opinions on different aspects of everyday life in popular social networking sites like Twitter, Facebook, and Tumblr. Due to the phenomenal growth of social networks, companies and media organizations are increasingly seeking ways to mine social media data and gain information regarding what people think about their products, services, and companies.

An online Social Network is a platform that allows users to create a profile and publish views and opinions. Users explicitly connect to other users thus creating social relationships. In a single sentence, we can say that social networks are user-generated content systems that allow their users to communicate and share information. There are different models followed by social networks like unilateral and bilateral. Twitter utilizes a social model of following which is unilateral, whereas Facebook uses a social model of friendship which is bilateral [1].

In classification algorithms, support vector machines are widely used to detect emotions of online users' text data. Abbasi et al [2] used SVMs in order to better understand the online user emotions and preferences. Alexander Pak [3] proposed a model which collects corpus from Twitter to perform opinion mining and sentiment analysis. He used SVM and Naïve Bayes machine learning algorithms to determine positive, negative, and neutral sentiments of a document. Most of the sentiment detection algorithms are designed to identify user opinions about products rather than user behaviors. Mike Thelwal [4] has implemented the SentiStrength algorithm to extract strength of sentiments from informal text. Xia Hu [5] proposed a sociological approach to handle noisy and short texts for sentiment classification. Nasir Naveed [6] has proposed a content based analysis of interestingness on Twitter using LDA and regression methods to show how tweets containing negative sentiment travel faster when compared to tweets with positive sentiment.

In Twitter, we have different features like "tweet", "retweet", "reply", "direct message", and "like". Out of all these features tweet is the principal feature through which users express their opinions, feelings, and reviews. Retweeting is a pattern that happens when another user has the same opinion or feeling that is expressed in the tweet, so, retweets diffuse the same information expressed by their original tweets. Like is a feature where users like the posts of their interest. Direct message is an option where messages are addressed to another user directly. These messages start with the username of the addressee, while other users can still see these messages, they are not the primary focus of the message [6].

We can see a lot of research work in sentiment analysis of tweets and retweets. In Twitter, texts are short due to limited number of characters, in contrast to customer reviews in e-commerce sites. To overcome this problem, we have designed a model which considers tweets and tweet replies in conjunction to analyze the sentiments of the tweets. Our goal is to build a model which can analyze sentiments of short texts by appending replies to the tweets using conjunctions.

Here is an example tweet:

Tweet: Fireworks on July 4th

Reply1: It's an eye feast

Reply2: Never like before, awesome

Reply3: Suffered with smog

Reply4: Nashville is hosting USA's biggest event

We can extract more information when we use reply along with the tweet text. In the above-mentioned tweet, tweet text alone is not expressing any positive sentiment or negative sentiment. If replies are considered using a conjunction with tweet text, it gives a better understanding of the data.

2. Data Description

A total of 21 gigabytes of data has been collected over 7 days in the JSON format. After analyzing the tweets, retweet patterns, and required features, we have selected some hashtags which have a significant amount of data. After this, we have parsed required data into a CSV file. The collected data has more than 30 hashtags, but we have selected 7 hashtags from different genres. We have collected data related to movies, web series, products, politically influential people, and trending events. The hashtags are: ElectionDay, FridayFeeling, ImsoOldSchoolThat, LivePD, Trump, USA, Russia, India, IphoneX, IoT, ThorRagnarok, Terrorists, ITMovie, and Havana.

We have collected the tweetId, fullText and textType from JSON data into a CSV file. Each hashtag data is collected in a different CSV file. In each file, we have tweets followed by replies. After collecting the data, we have labeled the sentiments of the data manually.

Twitter's developer platform offers several APIs and utilities to collect the required data like tweets and retweets using different inputs like tweetId and username. Twitter has two APIs: REST API and Streaming API. We have used REST API to collect the data. These APIs have limits to how many calls we can make a day. We have experimented with many API functions, but no function is providing us the data that we are looking for, so we have collected data based on memes (#hashtags). Out of that huge amount of data, we have collected the required data by parsing the JSON files. The required data is collected in CSV files, and features are extracted out of the parsed data.

2.1 Data Labeling

To implement the above mentioned methodology, we have labeled the tweets and tweet replies with positive and negative sentiments. We have labeled positive sentiments +1 and negative sentiments -1. Tweet replies are labeled with respect to the tweet sentiments. If reply is supporting the tweet, it is labeled with the same sentient value that the tweet is labeled with. If the reply is opposing the tweet then it is labeled with the opposite sentiment value that the tweet is labeled with. Data of each topic (hashtag) is available in a separate CSV file with sentiment values labeled.

2.2 Proposed System

The proposed system is shown in Figure 1.

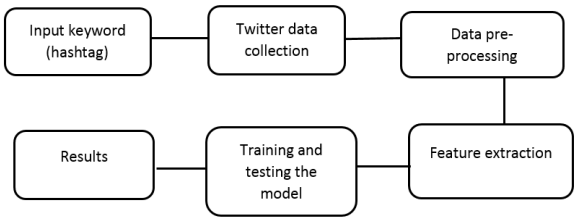

Figure 1: Proposed System

2.3 Hand-Crafted Features

The corpus collected from Twitter is huge and totaled 21 gigabytes of data. We have selected some required features like tweetId, textType, fullText. Based upon the few important features, we have hand crafted other features like direct message, includes username, includes hashtag, includes URL, parts of speech, positive terms, negative terms, positive emoticons, negative emoticons, exclamations, and question marks. All the hand-crafted features are content based features.

3. Proposed Method of Data Handling

In the process of finding a better way to analyze the sentiments of tweets, we have found replies of the tweets provide useful information which helps to analyze sentiments accurately. We can overcome the problem of analyzing sentiments from short texts by relying on the replies below the tweet. The tweet size is limited to only 140 characters, while their replies are sometimes less than 50 characters (very short text). The sentiment analysis algorithms and libraries are extremely effective for long texts but cannot produce a very good result for short text, which we have observed in the previous research work, An Approach for Pattern Recognition and Prediction of Information Diffusion Model on Twitter

To achieve a better result, we have used the existing libraries like Python Natural Language Tool Kit (NLTK) to analyze the problem. Our proposal is mainly to design a better model as shown in Figure 2. We have appended replies with their original tweet, to serve two purposes: first, as the number of replies increases, the size of text also increases which is suitable to determine the sentiment using NLTK [7], second, the sentiment of each of the replies can be evaluated in the context of its parent tweet. Primarily these two factors were missing in our previous research.

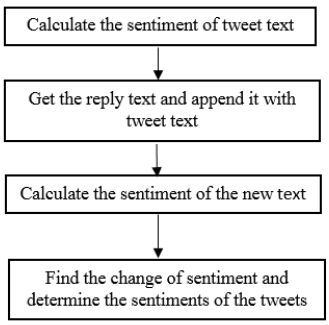

Figure 2: Proposed Method for Data Handling

4. Algorithms and Results

We have analyzed the datasets with appending replies and without appending replies. In the first step, NLTK is used to analyze the sentiments; it provides positive, negative and compound values of sentiments. NLTK gave a very low accuracy in both the scenarios (with and without appending tweets and replies). In the second step, we have planned to implement Support Vector Machine (SVM), Random Forest (RF), Decision Tree, and Naïve Bayes algorithms on the dataset with hand crafted features [8]. The accuracy of the algorithms on the merged dataset is high in both SVM and Random Forest when compared to the original dataset. In the merged dataset SVM outperformed the results of Random Forest as shown in Table 2, whereas Random Forest and SVM performed equally well on original dataset as shown in Table 1.

	Original Data			
Hasthags	**Naïve Bayes**	**Decision Tree**	**SVM**	**Random Forerst**
USA	0.651	0.680	0.706	0.702
ITMovie	0.268	0.707	0.975	0.926
LivePD	0.558	0.652	0.694	0.688
Russia	0.573	0.621	0.679	0.673
Terrorists	0.540	0.620	0.637	0.629
ThorRagna	0.637	0.646	0.681	0.673
Trump	0.622	0.655	0.707	0.689
AVG:	**0.551**	**0.660**	**0.725**	**0.711**

Table 1: Accuracy on Original Data

	Merged Data			
Hasthags	**Naïve Bayes**	**Decision Tree**	**SVM**	**Random Forerst**
USA	0.520	0.690	0.770	0.750
ITMovie	0.533	0.778	0.827	0.800
LivePD	0.632	0.760	0.810	0.775
Russia	0.567	0.694	0.760	0.739
Terrorists	0.678	0.771	0.754	0.752
ThorRagnarok	0.590	0.690	0.770	0.752
Trump	0.622	0.737	0.785	0.772
AVG:	**0.592**	**0.731**	**0.782**	**0.763**

Table 2: Accuracy on Merged Data

5. Contribution and Conclusion

In the current research work we have contributed a new data set and a new technique for analyzing the tweets. We found this new approach of handling data is successful at analyzing the short ambiguous texts on Twitter. We have used NLTK on the dataset and accuracy is recorded very low. In future work, we want to improve the accuracy of NLTK. Another improvement we can possibly make is topic modelling using Latent Dirichlet Allocation (LDA) [6].

6. References

[1] Guille, Adrien, et al. "Information diffusion in online social networks: A survey." *ACM Sigmod Record* 42.2 (2013): 17-28.

[2] Abbasi, Ahmed, Hsinchun Chen, and Arab Salem. "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums." *ACM Transactions on Information Systems (TOIS)* 26.3 (2008): 12.

[3] Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." *LREc*. Vol. 10. No. 2010. 2010.

[4] Thelwall, Mike, et al. "Sentiment strength detection in short informal text." *Journal of the Association for Information Science and Technology* 61.12 (2010): 2544-2558.

[5] Hu, Xia, et al. "Exploiting social relations for sentiment analysis in microblogging." *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013

[6] Naveed, Nasir, et al. "Bad news travel fast: A content-based analysis of interestingness on twitter." *Proceedings of the 3rd International Web Science Conference*. ACM, 2011.

[7] Perkins, Jacob. Python text processing with NLTK 2.0 cookbook. Packt Publishing Ltd, 2010.

[8] Bakliwal, Akshat, et al. "Sentiment analysis of political tweets: Towards an accurate classifier." Association for Computational Linguistics, 2013