Sentiment Analysis of Short Texts on Twitter Data

CSC 699

Fall 2017

Department of School of Computing

College of Science and Technology

THE UNIVERSITY OF
SOUTHERN
MISSISSIPPI

Committee Members

Dr. Andrew H. Sung

Dr. Zhaoxian Zhou

Dr. Kala R. Marapareddy

By

Asheshbabu Pothuraju – w982732

Declaration

I hereby declare that the Report of the Master Degree Project Work entitled Sentiment Analysis of Short Text on Twitter Data which is being submitted to the University of Southern Mississippi, Hattiesburg in partial fulfillment of the requirements for the award of the Degree of Master of Science in Computer Science in the School of Computing, is a bonafide report of the work carried out by me. The material contained in this Report has not been submitted to any University or Institution for the award of any degree.

Asheshbabu Pothuraju
w982732
School of Computing

Place: USM, Hattiesburg
Date: 11/30/2017

Acknowledgements

The successful completion of my project gives me an opportunity to convey my gratitude to each and every one who has been instrumental in shaping up the final outcome of this project. To begin with, I would like to express my sincere gratitude towards my advisor Dr. Andrew H. Sung (Director, School of Computing, USM, Hattiesburg), for allowing me to carry out the major project work under his guidance. His consistent support and positive attitude enabled me to complete this project successfully. I wish to express my sincere thanks to Amartya. H and Trung Nguyen who has been instrumental in the successful completion of this project. I express my gratitude towards my parents and my family members for the timely care and support extended to me, without which this project would not have been possible at all.

Asheshbabu Pothuraju

Abstract

Online social networks like Twitter, Facebook plays an integral role in everyday life of the people. Through this medium we can share information with our friends, acquaintances. Twitter is a platform where we can post our views, opinions and feelings. Twitter has become a digital media and marketing platform. Influential people like politicians, film starts post their views on Twitter. Products reviews, movie reviews are post on Twitter, which ultimately leads to marketing the products and films.

In this project, I have worked on 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 extracted different features from more than 20 million tweets to design a model to 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 handcrafted features. In the next step, different machine learning algorithms are used to train the model. SVM, Random Forest, NLTK, and LSTM are used to build the model. Random Forest outperformed other algorithms. 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, SVM, Random Forest, replies, NLTK, short texts.

Contents

1 Introduction

Microblogging today has become a very popular communication tool among Internet users. Millions of users share opinions on different aspects of everyday life in popular social networks like Twitter, Facebook, and Tumbler. Due to this phenomenal growth of social networks, companies, media organizations, and government organizations are increasingly seeking ways to mine these social l media for information about what people think about their products, services, and companies. Political parties are interested to know if people support their program or not.

A lot of data can be mined from online social networks, as the users post their opinions on many aspects of their everyday life. In this work, we present a method which performs unique method to calculate the sentiments of the Twitter data. We can see a lot of research work on sentiment analysis of Tweets, a bit less research work on retweets, but there is no research work on the sentiment analysis of the tweet replies. We have used different methods to leverage the features of the tweet replies to build a model which achieves a higher accuracy.

With the unprecedented increase in usage of internet and web technologies, number of people expressing their views, ideas and opinions on social networks are increasing. This online data is very useful for businesses, governments. Twitter produces around 6000 tweets a second on average, which counts to 500 million tweets per day, which is a huge source of information. Tweets have a limit of 140 characters. Twitter recently announce in September 2017, it was testing longer tweets. Some users praised the change while others feared the site would lose its sense of brevity. After that Twitter doubled the character limit of its tweets to 280 characters.

2 Literature survey

The first phase of the project is literature survey. A lot of previous work related to sentiment analysis on Twitter and Online Social Networks (OSN) have been studied and mining Twitter data and extracting sentiments and information diffusion pattern has been considered as the area of interest; other than that methods of handling skewed datasets have been studied. Here is a brief description about some of the related work in Twitter data mining, Information diffusion, and skewed data.

2.1 Survey of existing work

Online Social Network is site 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 a user-generated content system that permits its users to communicate and share information.

There are different models followed by social networks like unilateral and bilateral. Twitter follows a social model of following which is unilateral. Whereas Facebook follows a social model of friendship which is bilateral [1]. Twitter was created in 2006, this service took more than 2 years to gain worldwide popularity. By 2012 Twitter has more than 100 million users with millions of tweets a day. The tweets count, the amount of data produced by Twitter has increased rapidly throughout the world.

Researchers have started working on Twitter to make use of huge amount of data. AlSumait et al. [2] propose an online topic model, more precisely, a non-Markov on-line LDA Gibbs sampler topic model, called OLDA. Nasir Naveed [3] has proposed a content based analysis of interestingness on Twitter using LDA and regression methods to say bad news travel fast.

Adrien Guille [4] has performed a survey on information diffusion in online social networks and finally proposed a taxonomy and main research challenges arising from information diffusion in online social networks. Alexander Pak [5] proposed a model which collects corpus from Twitter to perform opinion mining and sentiment analysis. He has use SVM and Naïve Byes machine learning algorithms to determine positive, negative, and neutral sentiments of a document. A Tumasjan [6] implemented a model which predicts election results with Twitter data.

3 Proposed system

The project is divided in to 6 different phases as shown in the Figure 1.

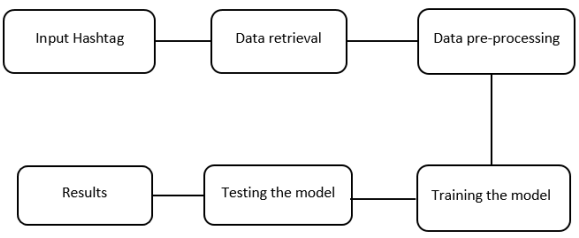

Figure 1: Proposed System

4 Insights of Twitter for Better Understanding

Twitter provides many features on its web and mobile platforms. Some of the features are tweet, retweet, reply, like, direct message, and follow. We can follow influential people like politicians, actors, CEOs of big companies. By doing this we will get to know more about the people whom we follow.

4.1 Direct Message

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 in primary focus of the message, shown in Figure 2 and Figure 3.

Figure 2: Direct message option

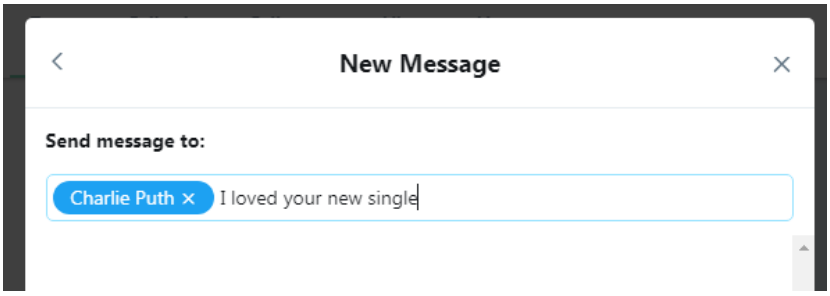

Figure 3: Direct message

4.2 Retweets

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 their original tweets,

shown in Figure 4 and Figure 5. When we add any additional text while retweeting, retweet converts in to a tweet, shown in figure 6.

Figure 4: Retweet option

Figure 5: Retweet window

Figure 6: Retweet turns into a tweet

4.3 Tweet Reply

These are the texts which are posted by different users in response to the tweet. Users will react to the situation, shown in Figure 8 and Figure 9.

Figure 7: Reply option

Figure 8: Reply window

Figure 9: Posted reply

5 Data description

Twitter is an online social network, which provides a facility to scrape data from it, unlike Facebook and other platforms. There are many APIs available for collecting the Twitter data. We can collect streaming data too using Twitter APIs [7].

We have used Twitter API to scrape the data Twitter uses a NoSQL database dude to huge amount of data that streaming every day. Relational databases cannot handle huge amount in the form of tables.

Python is used to collect data from Twitter. We have collected data using hashtags. Our input is a CSV file with the hashtag titles, and our output is a JSON data file with all the tweets, replies, and retweets which has the input hashtag title. Hashtags are basically the summarizer of the tweet and hence are very critical to collect the relevant information from hashtags.

Total 21 GB of data is collected in the form .JSON format. After analyzing the data patterns and required features, we have selected some hashtags which have more data. After this we have parsed required data in to a CSV file. The collected data has more than 30 hashtags, but we have selected 15 hashtags out of that from different genres. We have collected data related to movies, web series, products, politically influential people, and trending events.

- ElectionDay
- FridayFeeling
- ImsoOldSchoolThat
- Trump
- USA
- Russia
- India
- IphoneX
- IoT
- ThorRagnarok
- Terrorists
- ITMovie
- Havana

5.1 Data Selection

We have collected data from different genres like music, products, influential people, countries, social problems. Brief explanation of each hashtag is provided below.

- #ElectionDay talks about Election Day debates
- #FridayFeeling is about Friday parties
- #ImsoOldSchoolThat is a political discussion
- #LivePD is a television show
- #Trump is about president of USA
- #USA is about political discussion and happening things in USA
- #India is about political discussion and happening things in USA
- #Russia is about political discussion and happening things in USA
- #IphoneX is about reviews on Apple's new phone
- #IoT is about recent hacks
- #ThorRagnarok is a movie
- #Terrorists is about antisocial activities by Islamic state terrorists
- #ITMovie is a movie
- #Havana is a music album by Camila Cabello

We have collect the tweet id, tweet text and reply text from JSON data in to 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. Tweet's text is classified as positive (+1) and negative (-1). If the reply supports the tweet, we label the same sentiment the tweet is labeled with, if the reply opposes the tweet we label the sentiment which the tweet doesn't holds.

Total we have four columns tweet ID, type (which specifies the content is twee or reply), full text, and sentiment. We have more than 4000 records (which is sum of tweets + replies) in our dataset. It is a unique data set where we can find tweets with the replies followed by it.

6 Data Pre-processing

After collecting the data based on hashtags, we have processed data. Some of the important stages are explained briefly in the below sections.

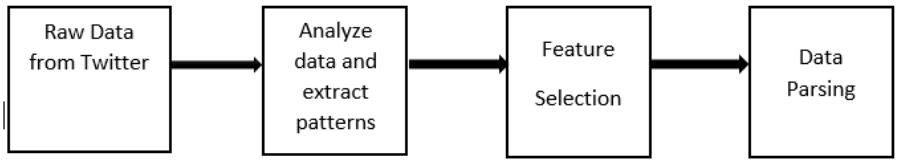

Figure 10: Different stages in data pre-processing

6.1 Tokenization

After downloading the tweets, replies, retweets using the hashtag provided in the on hashtag, we first tokenize the tweets.

6.2 Removal of non-English words

Twitter supports more than 30 languages. However, this work deals only with the English tweets. We have collected English tweets with help of language variable from the data [7].

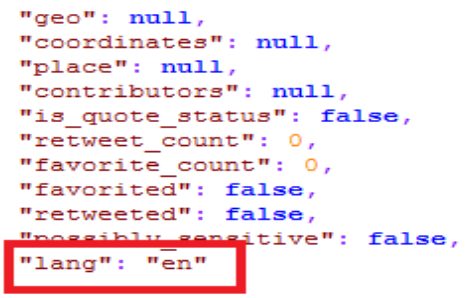

Figure 11: Language attribute in JSON data

Name	Language code
English (default)	en
Arabic	ar
Bengali	bn
Czech	cs
Danish	da
German	de
Greek	el
Spanish	es
Persian	fa

Figure 12: Languages supported by twitter

6.3 Natural Language Tool Kit (NLTK)

Natural Language Toolkit is a Natural Language Processing framework. Natural language we mean is a language that is used for everyday communications by humans; languages like English, Arabic or Portuguese. Natural Language processing, NLP for short involves complete understanding of human utterances, at least to the extent of being able to give proper responses to them.

NLTK was created in 2001 as part of a computational linguistics course in the Department of Computer and Information Science at the University of Pennsylvania. Since then it has been developed and expanded to many languages with the help of dozens of contributors. Table below lists the most important modules of NLTK [8].

Language processing task	NLTK modules	Functionality
Accessing corpora	nltk.corpus	Standardized interfaces to corpora and lexicons
String processing	nltk.tokenize, nltk.stem	Tokenizers, sentence tokenizers, stemmers
Collocation discovery	nltk.collocations	t-test, chi-squared, point-wise mutual information
Part-of-speech tagging	nltk.tag	n-gram, backoff, Brill, HMM, TnT
Classification	nltk.classify, nltk.cluster	Decision tree, maximum entropy, naive Bayes, EM, k-means
Chunking	nltk.chunk	Regular expression, n-gram, named entity
Parsing	nltk.parse	Chart, feature-based, unification, probabilistic, dependency
Semantic interpretation	nltk.sem, nltk.inference	Lambda calculus, first-order logic, model checking
Evaluation metrics	nltk.metrics	Precision, recall, agreement coefficients
Probability and estimation	nltk.probability	Frequency distributions, smoothed probability distributions
Applications	nltk.app, nltk.chat	Graphical concordancer, parsers, WordNet browser, chatbots

Figure 13: NLTK features

NLTK was designed with four primary goals in mind:

- Simplicity
- Consistency
- Extensibility
- Modularity

Natural Language Processing typically uses large bodies of linguistic data, or corpora. Corpus is a large body of text. Many corpora are designed to contain a careful balance of materials in one or more genres. Gutenberg Corpus is a text corpus, which contains 25,000 free electronic books.

Python is a simple yet powerful programming language with excellent functionality for processing linguistic data [8]. NLTK is an open source python library which provides an infrastructure that can be used to build NLP programs in Python.

7 Methodology

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 very effective for long texts but cannot produce a very good result for short text, which we have observed in the previous research work, Information Diffusion on Twitter: Pattern Recognition and Prediction of Volume, Sentiment, and Influence.

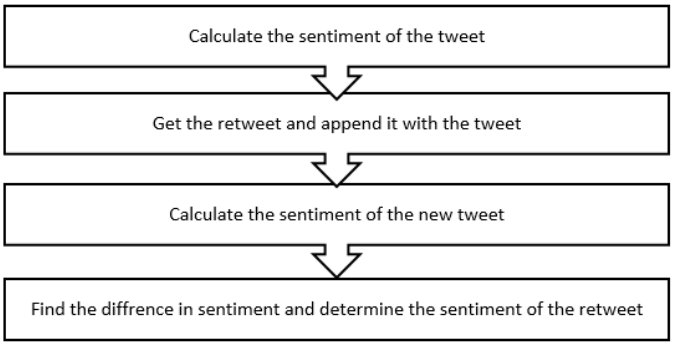

Figure 14: Data handling technique

To achieve a better result, we have used the existing libraries like Python NLTK to analyze the problem. Our proposal is mainly to design a better model as shown in Figure 1. We have appended replies with their original tweet, to serve two purposes: First, as the number of replies increases, the size of text is also increases which is suitable to determine the sentiment using NLTK.

Secondly, the sentiment of each of the replies can be can be evaluated in the context of its parent tweet. Primarily these two factors were missing in our previous research.

8 Hand Crafted Features

After analyzing the data that we have collected from Twitter, we have created a data set with required features. Features are

- Direct message
- Includes Username
- Includes Hashtag
- Includes URL
- Exclamation mark
- Question mark
- Term Positive
- Term negative
- Emoticon positive
- Emoticon Negative
- Parts of speech

9 Algorithms Implemented

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

A Random Forest algorithm takes the decision tree concept further by producing a large number of decision trees. The approach first takes a random sample of the data and identifies a key set of features to grow each decision tree. These decision trees then have their out of bag error determined (error rate of the model) and then the collection of decision trees are compared to find the joint set of variables that produce the strongest classification model.

The most common application of decision trees is within classification. The manner in which we evaluate the performance of classification algorithms is distinctively different from continuous models (like regression analysis). Because we know the correct class of the data, one common metric that is used is the overall predictive accuracy. However, this approach can sometimes be problematic when the classes are not balanced.

Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well [9].

A categorical variable can be represented in a cross tab (TN & TP) which allows us to see how well a model had performed. The left side of the matrix shows the actual scenario based on historical data and the top shows the predicted results of the model that we are evaluating. Short notations are handy to represent like, True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) rates of the model.

In order to perform an ROC analysis, we need to calculate some figures from the confusion matrix. It is highly unlikely that we will create perfect predictive models from data that is available with us. There will be misclassifications and prediction errors which have to be considered in the performance evaluation. Specificity and sensitivity are statistical measures of the performance of a binary classification and these metrics are a critical components to the ROC analysis and they represent aspects of the modeling error. Specificity is the probability that the model will predict negative results. Sensitivity is the probability that the model will predict even.

10 Results and Discussion

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. The accuracy of the algorithms on the merged dataset is high in both SVM and RF when compared to the original dataset. In the merged dataset SVM outperformed the result of Random Forest whereas on the original dataset Random Forest and SVM performed equally well Table.1 and Table 2.

	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

I have compared average values of SVM, RF, Decision Tree, and Random Forest in two different scenarios, shown in Table 3.

	Original Dataset	Merged Dataset
Naïve Bayes	0.551	0.592
Decision Tree	0.660	0.731
SVM	0.725	0.782
Random Forerst	0.711	0.763

Table 3: Comparison of average values

11 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 very good at analyzing the short ambiguous tweets on Twitter. We have used LSTM and NLTK on the dataset. Accuracy of LSTM and NLTK are very poor. In the future work we want to implement LSTM, NLTK in a better way. Another improvement we can possibly make is topic modelling on the corpus using Latent Dirichlet Allocation (LDA) [3].

12 References

[1] Guille, A., Hacid, H., Favre, C., & Zighed, D. A. (2013). Information diffusion in online social networks: A survey. *ACM Sigmod Record*, *42*(2), 17-28.

[2] AlSumait, L., Barbará, D., Gentle, J., & Domeniconi, C. (2009). Topic significance ranking of LDA generative models. *Machine Learning and Knowledge Discovery in Databases*, 67-82.

[3] Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011, June). Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd*

[4] Guille, A., Hacid, H., Favre, C., & Zighed, D. A. (2013). Information diffusion in online social networks: A survey. *ACM Sigmod Record*, *42*(2), 17-28.

[5] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010).

[6] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm*, *10*(1), 178-185.

[7] https://dev.twitter.com/web/overview/languages

[8] Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc

[9] Gupta, M., Dalmia, A., Jaiswal, A., & Reddy, C. T. Sentiment Analysis in Twitter. *International Web Science Conference* (p. 8). ACM.