TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 6557 - 6560

Twitter Data Preprocessing for Sentimental Analysis

[1]P. Amaresh, [2]M. Raja Suguna

[1]UG Scholar, [2]Assistant Professor,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences, Chennai
[1]peruguamaresh28@gmail.com,[2]suguna.raj89@gmail.com

Article Info
Volume 82
Page Number: 6557 - 6560
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 01 February 2020*

Abstract
With the innovation of web and its development, there is a large amount of information present in the web for web clients and a great deal of information is produced daily as well. Informal communication destinations like Twitter, Facebook, Google+ are quickly picking up prominence as they enable individuals to share and express their views, post messages over the world. Lot of work has been done to find the Sentiment of the Tweets. Identifying the sentiments behind the tweets helps in Business, where marketing companies develop strategies by recording the nature as well as quantify the respond by the customer during new product launch. Also useful in politics, where people view can be tracked, understanding the consistency of the statements at the government level, even used to predict the results of the election. In this paper, we give a study and a near investigations of existing methods for assessment mining like AI and vocabulary-based methodologies, together with assessment measurements. Utilizing different machine learning algorithms like Naive Bayes, Max Entropy, and Support Vector Machine, we give polarity on twitter information.

Keywords: *Twitter, Facebook, Web Information, AI, Naive Bayes, Support Vector Machine.*

1. Introduction

Now-a-days, different social media and streaming platforms like Twitter, Facebook, Myspace, YouTube, Netflix have picked up so much prominence and we can't disregard them. They have gotten one of the most significant utilizations of Web. They enable individuals to share picture sharing, online journals, wikis and so forth.

Sentimental Analysis Also referred as Opining Mining which use the Natural Language Processing Techniques to reveal the emotion, opinion of a piece of writing such ad tweets, Review in E-Commerce Websites and You tube videos. It also referred as predicting the Polarity of the writings. The Polarity refers to the Positive Statement or Negative Statement or a Neutral Statement.

Natural Language process, sometimes shortened as information processing, is a branch of AI that deals with the interaction between computers and humans victimization the tongue. The objective of information processing is to browse, decipher, understand, and add up of the human languages in a very manner that's valuable. Most information processing techniques relay on machine learning to derive meaning from human languages.

With the recent advances in deep learning, the flexibility of algorithms to analyse text has improved significantly. Artistic use of advanced AI techniques may be an efficient tool for doing in-depth analysis. An extension or data model that detects the hate content from the twitter data and perform sentimental analysis and keyword detection report so that millions of tweets are read and analyzed to stop the wide spread of hatred and ill social behavior. This system is able to provide a good understanding about what views people want to express and give a complete summary of public opinion thus by

January-February 2020
ISSN: 0193-4120 Page No. 6557 - 6560

opening new gates to improve quality of the products for business etc.

Main objective of this project is to perform sentimental analysis on the tweets posted by the people in twitter and supply a report on the particular parameters.

2. Related Works

Yuan et al [1] used deep learning models Recurrent Neural Network (RNN) and Recurrent Neural Tensor Net (RNTN) and two hidden layers RNN to train the preprocessed Tweets. The Preprocessed tweets are represented as a binary dependence tree. They tuned their network with several hyper parameters, but the model results in a poor outcome for prediction of negative tweets.

D. Mahajan et al [2] used Recurrent Neural Network to train their preprocessed tweets. They also make use of Google Translator to translate the sentences in other languages to English for the effective prediction of sentiment of the tweets. Comparisons with the other Machine Learning Algorithm such as Support Vector Machines and Naïve Bayes method, RNN shows the more accuracy in predicting the polarity of the tweets.

Y. M. Wazery et al [3] used Recurrent Neural Network along with Long Short Term Memory (LSTM) library to trained their preprocessed datasets taken from Amazon, airline and IMDB sites and compared their work with machine Learning algorithm such as K-nearest Neighbour algorithm, Decision tree, Support Vector Machines and Naïve -Bayes algorithm. The deep learning method RNN with LSTM method achieves more accuracy than other machine Learning Algorithms.

Liaoet et al [4] proposed convolutional Neural Network to train the preprocessed review and Gold Dataset containing tweets.CNN got more accuracy in predicting the polarity of the tweet than SVM and Naïve Bayes Algorithms.

3. System Architecture

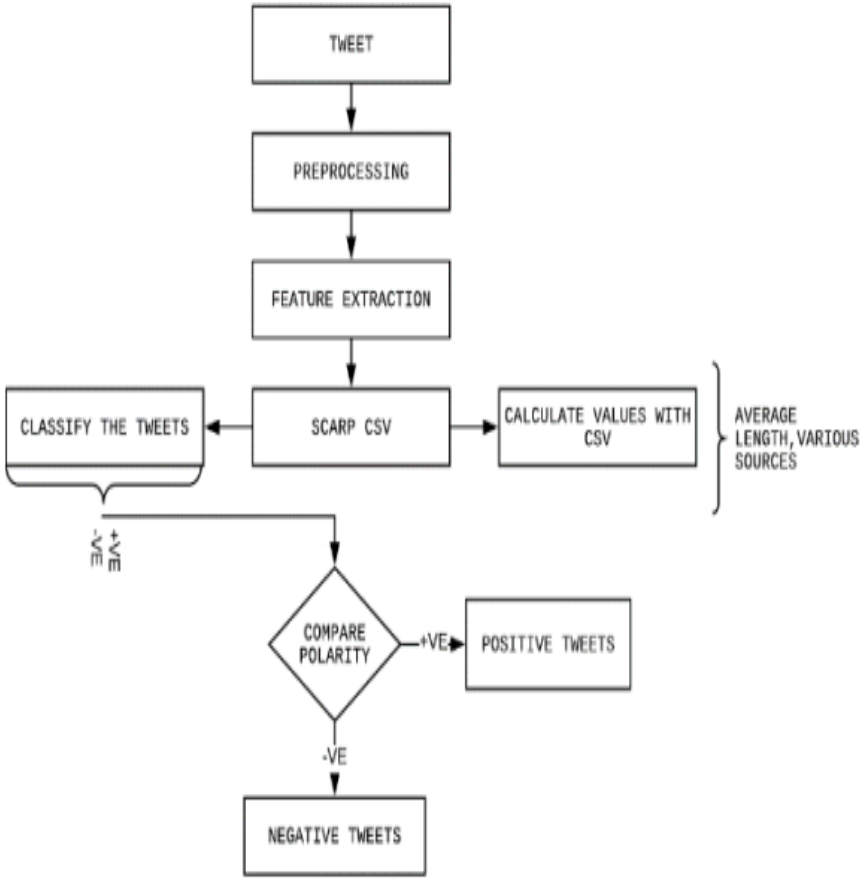

Figure 1: System Architecture

4. Tweet Preprocessing

In order to predict the polarity oh the tweet, first need to download the Tweets from the Twitter. For this Purpose, one should have a Tweet Account in this Micro Blogging Site called Twitter. Then Need to get Twitter API Connection using Twitter Development Site of Twitter Account. Twitter API Connection provide the users with Credentials such as API Key and API Token

TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 6557 - 6560

for Downloading the Tweets from the Twitter Development Site.

4.1 Removing White Spaces, Punctuations, Stop Words, Tweet Handles, Numbers Special Characters

The tweet is a short message. It contains piece of sentences having white spaces, punctuations and so on. Words will contribute for finding the sentiment behind them. Other factors such as punctuations and whitespaces, emojis will doesn't refers any sentiments. So all the white Spaces, Punctuations, Twitter handles (@User), Numbers, Special Characters and Stop Words are removed from each Tweet. Stop Words Refers to the commonly used words in a language. The English contains Stop Word List of 32 Words. ex: and, or, across, before, after etc. By removing the Stop Words, can concentrate only on Important Words. Fig 1. Shows how the Raw Tweet and the Tidy Tweets obtained after Preprocessing.

	id	label	tweet	tidy_tweet
0	1	0.0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run	when father dysfunctional selfish drags kids into dysfunction #run
1	2	0.0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked	thanks #lyft credit cause they offer wheelchair vans #disapointed #getthanked
2	3	0.0	bihday your majesty	bihday your majesty
3	4	0.0	#model i love u take with u all the time in urð□□±!!! ð□□□ð□□□ð□□□ð□□□ ð□□¦ð□□¦ð□□¦	#model love take with time
4	5	0.0	factsguide: society now #motivation	factsguide society #motivation

Figure 2: Raw tweets and Tidy Tweets (Cleaned Tweets)

4.2 Tokenization

Tokenization is the process of breaking the Strings in Tweet into individual words called Tokens, Fig 2 shows the Tokens.

```
0                [when, father, dysfunctional, selfish, drags, kids, into, dysfunction, #run]
1    [thanks, #lyft, credit, cause, they, offer, wheelchair, vans, #disapointed, #getthanked]
2                                                                  [bihday, your, majesty]
3                                                        [#model, love, take, with, time]
4                                                     [factsguide, society, #motivation]
Name: tidy_tweet, dtype: object
```

Figure 3: Tweet Tokens

4.3 Stemming and Lemmatization

In order to provide different grammatical meaning,a single word can be inflected (modified) at prefix, suffix or an internal modification such as vowel Change. Stemming is Process of removing the suffixes from the words by using a Rule Based Approach. Thus the root word is identified for three or more inflected words and replaced by the root word. Table 1. Shows the Stemming Process.

Table 1: Stemming

Form	Suffix	Stem
studies	-es	studi
studying	-ing	study

Lemmatization also stripes out the inflections from the modified word but by considering the language morphological meaning. For this technique a detailed dictionaries for the respective language is needed by the lemmatization algorithm to find out the original root word and its corresponding morphological meaning. Table 2, gives the lemmatization example.

Table 2: Lemmatization

Form	Morphological information	Lemma
studies	Third person, singular number, present tense of the verb study	study
studying	Gerund of the verb study	study

All the above preprocessing techniques gives cleaned tweet as a output. After Preprocessing of tweets, it can be visualized by using Word Cloud and Bar graph. Word Cloud is a Visualization technique

January-February 2020
ISSN: 0193-4120 Page No. 6557 - 6560

where most frequent letter shown as larger one and less frequent words are displayed as smaller in size.

5. Results and Discussion

After Preprocessing, features need to be extracted from the tweets. These features are extracted by methods such as Bag of Words, TF-IDF measure. In Bag of words, a Corpous is Created which contains two attributes D and F. D refers to the total number of documents and F refers to the frequency of each term in the document. Thus, Bag of Words is a matrix of order "D × F". TF-IDF measure refers to Term Frequency Measure and Inverse Document Frequency, that is besides considering all the frequency of terms in the document, it also consider the importance of words having lower frequency. After a Model need to be trained by these features to predict the polarity of the tweets.

6. Conclusion

A Detailed literature survey has been carried on Twitter Sentimental Analysis. Tweets have been downloaded from the twitter social Networking site using Twitter API Connection. All tweets have been preprocessed by using Tokenization and Bag of Words is Created. Stemming and Lemmatization techniques applied on the tweets to correct it to phrases. These preprocessed tweets are modeled by using Data Mining Algorithms to reveal the sentiment behind each tweet.

References

[1] Yuan, Ye and You Zhou. "Twitter Sentiment Analysis with Recursive Neural Networks." (2015).

[2] D. Mahajan and D. Kumar Chaudhary, "Sentiment Analysis Using Rnn and Google Translator," 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, 2018, pp. 798-802.

[3] Y. M. Wazery, H. S. Mohammed and E. H. Houssein, "Twitter Sentiment Analysis using Deep Neural Network," 2018 14th International Computer Engineering Conference (ICENCO), Cairo, Egypt, 2018, pp. 177-182.

[4] Liao, Shiyang, Junbo Wang, Ruiyun Yu, Koichi Sato, and Zixue Cheng, "CNN for situations understanding based on sentiment analysis of twitter data", Procedia Computer Science, vol. 111, pp. 376-381, 2017.

[5] Jianqiang, Zhao, GuiXiaolin, and Zhang Xuejun, "Deep convolution neural networks for twitter sentiment analysis", IEEE Access, vol. 6, pp. 23253-23260, 2018