International Journal of Research in Engineering, Science and Management
Volume-3, Issue-5, May-2020
www.ijresm.com | ISSN (Online): 2581-5792

Analyzing Sentiments in Twitter Using Machine Learning

Pooja S. Nair[1*], P. Parvathi[2], V. Shahina[3], B. Mridula Parthan[4], K. P. Biji[5]
[1,2,3,4]*Student, Department of Computer Science and Engineering, Nehru College of Engineering and Research Centre, Thrissur, India*
[5]*Assistant Professor, Department of Computer Science and Engineering, Nehru College of Engineering and Research Centre, Thrissur, India*
*Corresponding author: poojanairpsn@gmail.com

***Abstract*: Twitter stands not only as a social media but also to easily understand and categorize, reveal customers' needs, mindsets and gather information regarding any campaign or interests. Analyzing the sentiments of tweets help us to study the polarity of thoughts of certain people or product. Through this obtained data, any tweets posted by a user in twitter can be indicated using its polarity like positive, negative and unbiased tweets. This enhanced sentimental analysis of twitter has retrieved the data based on an inception and performs using Text Blob. The proposed system is to examine the pre-stored data by real-time analysis via Twitter API. As a result, it helps to look over the posts with better accuracy using machine learning.**

***Keywords*: Text Blob (library for processing textual data), Data mining, Micro blogging.**

1. Introduction

Twitter, a much popular social media which connects people, thereby giving their opinions on different trending topics. At present times, people are exposing their social and private life through social media using tweets, reviews, hashtags comments, posts and emojis.

[2] This application has been developed as a significant smaller scale blogging site, having more than 100 million clients producing more than 500 million tweets per day [1]. Twitter clients are permitted to impart their insights as tweets, utilizing just 140 characters. This prompts individuals to compact their announcements by utilizing slang shortenings, emojis, short structures and so forth.

Moreover, this platform is also being used as a marketing strategy by businesses to connect with their customers. Mainly, this marketing strategy is to learn more about the customers, to make buzz about the new product, to obtain a very fast customer review and mostly to brand a much more human, Twitter gets us to covered on all these.

We explore the method for building such data using Twitter hashtags (e.g, #CoronaVirus, #Stayhome, #socializing) to identify positive, negative, and neutral tweets to use for training three-way sentiment classifiers. Thus, these tweets and the hashtags are must for analyzing the thinking level of individual people.

The Classifier will classify the tweets according the training set and regulates the polarity of the tweet as the output. The test data is basically the real-time tweets from twitter accessed using Twitter API. In this paper, we are going to analysis the microblog called as Twitter, classify the "tweets" into positive, negative and neutral sentiment.

2. Literature Survey

As the number of transactions in E-market places is growing drastically, more and more product information and reviews are available on the Internet. [3] Since costumers wants to purchase good products, reviews became most salient information. But due to massive quantity of reviews, customers can't consider all reviews. In order to resolve this issue, a lot of research is being conducted in Opining Mining. Through Opinion mining, contents of whole product reviews will be made available. Nowadays computational statistics are applied to handle massive volume of reviews. A method for summarization of product reviews using the user's opinion, feature occurrences and the rate of review in order to improve the performance of existing methods was proposed by Jung-Yeon Yang, Jaeseok Myung and Sang-goo-Lee. Through this method, enormous amount of reviews can be handled in a short span of time efficiently.

[4] Blog texts are classified according to the mood reported by its author during the writing. The data consists of a huge collection of blog posts – online diary entries – which include an indicator of the writer's mood. A system that Experiments with Mood Classification in Blog Posts by proposed by G Mishne.

The main finding was that mood classification was a demanding task using current text analysis methods. A diversity of features for the classification process was being used, which included content and non-content features, and some features which were distinctive to online text such as blogs. The results show a small, consistent, improvement over a naive baseline; while the success rates were relatively low, human performance on this task was not substantially better. So, this method was proved better than the traditional human performance.

Hate speech detection on Twitter is critically used in applications like event extraction, building AI chat bots, content recommendation systems, and sentiment analysis. This task is defined as being able to classify a tweet as racist, sexist or none. A system that detects Hate Speech Detection in tweets using deep learning was proposed by Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, Vasudeva Varma [5]. Complexity of the natural language constructs made this problem very challenging. Extensive experiments with multiple deep learning architectures were used to learn semantic word embeddings to handle this complex situation. This experiments on a benchmark dataset of 16K annotated tweets proved that these deep learning methods outperform state-of-the-art char/word n-gram methods by ~18 F1 points.

Vulgarity is a common linguistic expression and is used to perform several linguistic functions. Considering that most thoughts can easily be reworded to not include vulgar language; the use of explicit words indicates a strong desire to send a specific message. Expressively Vulgar: The Socio-dynamics of Vulgarity a paper proposed by Isabela Cachola, Eric Holgate, and Junyi Jessy Li, researchers at the University of Texas and University of Pennsylvania [6].

More specifically, their research analyzes the socio-cultural and practical aspects of vulgar language in tweets. In this study, the researchers compiled a dataset of 6,800 tweets. Next, they had the tweets labeled for sentiment by nine annotators, using a five-point scale. The data also includes demographics (gender, age, education, income, religious background, and political ideology) of those who posted the tweets. Further, w sentiment ratings for vulgar tweets were collected to study the relationship between the use of vulgar words and recognized sentiment and show that explicitly modeling vulgar words can boost sentiment analysis performance.

3. Methodology

As shown in the Fig 1.1. the system architecture consists of components such as Tweets extraction from twitter, where the data to be processed is gathered from twitter, and the preprocessing of data, feature extraction, Training set are defined for the given analysis. For the processing, the predefined set of positive or negative tweets is obtained from Kaggle dataset which is trained using Random forest classifier and output is the polarity of the searched tweet.

In this paper, we used Text Blob as a procedure to find the polarity of the text (positive, negative or neutral). The tweets are imported from the Twitter using API provided by the Twitter Developer on request. From these API various fields like tweets, source, retweets, likes, language, user etc. can be fragmented. On collecting these data, analysis of famous person thoughts on an event or occasion can be validated.

After creating a lot of custom features, utilizing bag-of-words and word2vec representations and applying random forest classifier, the classification accuracy level of 58% was achieved.

In order to create Twitter API, the user initially request for its authentication credentials, which will be of unique access tokens and keys. Hence, the tweets are retrieved based on the input keyword and the number of tweets to be counted for polarity check.

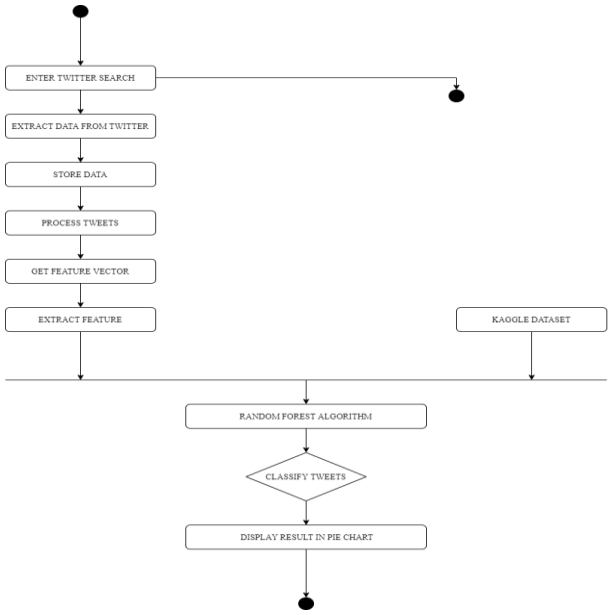

Fig. 1. System architecture

In the reprocessing of the data to create a bag-of-words representation of data. The steps mentioned below needed to executed as follows:

1. Cleansing
2. Remove URLs
3. Remove usernames (mentions)
4. Remove tweets with *Not Available* text
5. Remove special characters
6. Remove numbers
7. Text processing
8. Tokenize
9. Transform to lowercase
10. Stemming
11. Build word list for Bag-of-Words

```
In [13]: data = TwitterData_TokenStem(data)
         data.tokenize()
         data.stem()
         data.processed_data.head(5)
```

Out[13]:

	id	emotion	text	tokenized_text
1	6359301692413747 20	neutral	[io, app, transport, secur, mm, need, to, chec...	[IOS, App, Transport, Security, Mm, need, to, ...
2	6359502586825236 48	neutral	[mar, if, you, have, an, io, devic, you, shoul...	[Mar, if, you, have, an, iOS, device, you, sho...
3	6360308034330091 53	negative	[my, phone, doe, not, run, on, latest, io, whi...	[my, phone, does, not, run, on, latest, IOS, w...
4	6361009062248488 96	positive	[not, sure, how, to, start, your, public, on, ...	[Not, sure, how, to, start, your, publication,...
5	6361762729477447 72	neutral	[two, dollar, tuesday, is, here, with, forklif...	[Two, Dollar, Tuesday, is, here, with, Forklif...

Fig. 2. Preprocessed data

The processed data may be displayed as shown in Fig. 2. On Experimenting with the training set with polarity to apply the machine learning algorithms we check the accuracy of each algorithm.

A. *Bag-of-words+Naïve Bayes Algorithm*

Since the Bag-of-words representation being binary, we use the most common algorithm Naïve-Bayes Classifier which was based on train: test ration 7:3 stratified split.

The result is of 58% accuracy but this accuracy turns out to be different in final testing set where, 8-fold validation is used for optimizing 8-core machine which results in an accuracy of 43%.

B. *Random forest algorithm*

Considering the accuracy loss in Naïve-Bayes algorithm we used Random forest algorithm which provides with much better results while considering the full model, extended features, and the average accuracy from the cross validation is almost 58% which is much more stable than the Naïve-Bayes Algorithm. Hence the results of various experiments are shown in table 1.

Table 1
Accuracy of classifiers

Algorithm	Accuracy
Bag-of-words	40.032
Naïve-Bayes+ Bag-of-words	43.167
Naïve-Bayes+ KNN	54.876
KNN+ BOW	51.657
Random Forest+Bag-of-words	58.632

Once the algorithm Random Forest provides the best accuracy, the fields are extracted and segregated CSV is created. Using this data, the length of the message, Likes, retweets for the id is extracted and various results are derived. With the scraped tweets, classified into positive or negative or neutral.

4. Result

This is the output format for the project where we enter the key word as displayed below:

```
Enter Keyword/Tag to search about: Corona virus
Enter how many tweets to search: 300
How people are reacting on Corona virus by analyzing 300 tweets.

General Report:
Weakly Positive
```

Fig. 3. Keyword to search and count

```
General Report:
Weakly Positive

Detailed Report:
14.00% people thought it was positive
11.67% people thought it was weakly positive
4.33% people thought it was strongly positive
4.00% people thought it was negative
22.00% people thought it was weakly negative
0.67% people thought it was strongly negative
43.00% people thought it was neutral
```

Fig. 4. Detailed report

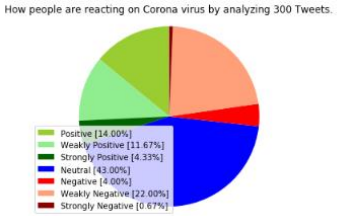

Fig. 5. Pie-chart

5. Conclusion

Sentiment analysis of tweets provides opportunities across many fields, from business to politics. Being able to analyze tweets in real-time, and determine the sentiment that underlies each message, adds a new dimension to social media monitoring. Twitter Sentiment Analysis allows a deeper understanding of people's feelings and opinions. It adds an extra layer to the traditional metrics used to analyze the performance of brands on social media and provides powerful opportunities for improvement. Moreover, the experiments we conducted roved that predicting the text sentiment using machine learning is a non-trivial. Since in every experiment we analyzed that simply bag-of-words is not enough and more additional features are necessary for creating a better algorithm. Word2vec representation significantly the predictions quality. Hence sentiment analysis has a very bright scope of development in future.

References

[1] Shobhana G, Vigneshwara B, Maniraj Sai A, "Twitter Sentimental Analysis", Issue 4, November 2018.
[2] M Y V Nagessh, T Anuradha," A word2Vector Representation for Twitter Sentimental Analysis", JOICS, Volume 9, Issue 11, 2019.
[3] Anjana Jimington, "A Baseline based deep learning approach of live tweets," IJIRT, Volume 6, Issue 1, June 2019.
[4] Jung-Yeon Yang, Jaeseok Myung and Sang-goo Lee, "The Method for a summarization of product Reviews using user's opinion" IEEE, 2009, International Conference.
[5] Gilad Mishne, "Experiments with Mood Classification in Blog Posts", 2005, Informatics Institute, University of Amsterdam.
[6] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta: "Deep Learning for Hate Speech detection in Tweets ", 2017.
[7] Eric Holgate, Daniel Preot¸iuc-Pietro, Junyi Jessy Li: "Expressively Vulgar: The socio-dynamics of vulgarity and its effects of sentimental analysis in social media", proceedings of 27th International conference of Computational Linguistics, pp. 2927-2938, Santa Fe, New Mexico, 2018.
[8] N. Oliveira, P. Cortez, and N. Areal, "The impact of microblogging data for stock market prediction: using twitter to predict returns, volatility, trading volume and survey sentiment indices," Expert Systems with Applications, vol. 73, pp. 125-144, 2017.
[9] Random Forest. http://www.stat.berkeley.edu/breiman/RandomForest/ cc home.html.
[10] M. Birjali, A. Beni-Hssane, and M. Erritali, \Analyzing social media through big data using info sphere big insights and apache ume," Procedia Computer Science, vol. 113, pp. 280-285, 2017.
[11] Y. Huang, D. Guo, A. Kasako, and J. Grieve, \Understanding us regional linguistic variation with twitter data analysis," Computers, Environment and Urban Systems, vol. 59, pp. 244-255, 2016.
[12] Jansen B. J, Zhang, M. Sobel, K. and Chowdury, A. (2009), "Twitter power: Tweets as electronic word of mouth", Journal of the American Society for Information Science and Technology 60(11):2169–2188.
[13] Nehal Mamgain, Ekta Mehta, Ankush Mittal and Gaurav Bhatt, "Sentiment Analysis of Top Colleges in India Using Twitter Data", (IEEE) ISBN -978-1-5090-0082-1, 2016.
[14] Varsha Sahayak, Vijaya Shete and Apashabi Pathan, "Sentiment Analysis on Twitter Data", IJIRAE, January 2015.
[15] David Zimbra, M. Ghiassi and Sean Lee, "Brand-Related Twitter Sentiment Analysis using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks", IEEE, pp. 1530-1605, 2016.