Sentiment Analysis with NLP on Twitter Data

Md. Rakibul Hasan[1], Maisha Maliha[2], M. Arifuzzaman[3]
[1−2]Department of Electronics and Communications Engineering, East West University, Dhaka-1212
[3]Bangladesh University of Engineering and Technology, Dhaka-1000
[1]hasanrakib373@gmail.com, [2]mahimaliha10@gmail.com, [3]arif@iict.buet.ac.bd

***Abstract*—Every social networking sites like facebook, twitter, instagram etc become one of the key sources of information. It is found that by extracting and analyzing data from social networking sites, a business entity can be benefited in their product marketing. Twitter is one of the most popular sites where people used to express their feelings and reviews for a particular product. In our work, we use twitter data to analyze public views towards a product. Firstly, we have developed a natural language processing (NLP) based pre-processed data framework to filter tweets. Secondly, we incorporate Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) model concept to analyze sentiment. This is an initiative to use BoW and TF-IDF are used together to precisely classify positive and negative tweets. We have found that by exploiting TF-IDF vectorizer, the accuracy of sentiment analysis can be substantially improved and simulation results show the efficiency of our proposed system. We achieved 85.25% accuracy in sentiment analysis using NLP technique.**

***Index Terms*—Natural language processing (NLP), Twitter, data mining, Sentiment analysis.**

I. Introduction

Now-a-days, internet services are generating a large amount of data which is increasing significantly day by day [1]. Social networking sites are being used for microblogging where it has become a tremendous tool among Internet users for communication [2]-[3]. Every big and small company are joining the social networking site to share their product and try to know the reviews of the products from the consumer. The company will use sentiment analysis to grasp the opinion of shoppers concerning their merchandise, so they will analyze client satisfaction and as per that they will improve their product. Particularly, the developed method to sentiment analysis using, by and large, to look at between any device, public figure, Sports team and so on.

Twitter is the second biggest social networking platform after Facebook which generates 347,222 tweets every each minute and 21 million tweets per hour [1]. So it creates an opportunity for data mining and sentiment analysis based on users tweet. Since sentiment analyses are part of the data mining that can observe public educe about various topics and products. It is also the stem of natural language processing, text analysis, computational linguistics, bio-metrics, machine learning methods. We are choosing Twitter for sentiment analysis because it offers opportunities for the tenderness of enunciated disposition. Twitter is limited to 140 characters of text that's why users can explain their brief ideas via a short message [4].

In this Paper, we have developed an NLP based pre-processed data framework to filter tweets where we incorporate Bag of Word (BoW) model and TF-IDF (Term Frequency - Inverse Document Frequency) model concept to sentiment analysis. In NLP Technique, it has done tokenization, stemming, lemmatization, removal of stopwords, POS tagging, named entity recognition, coreference resolution, and text modeling as Bag of Word and TF_IDF Model [5]. The main aim is to identify the sentiment of the tweet by defining positive and negative polarity where tweets are collected by using Twitter streaming API from Twitter. We utilize these tweets as crude data. At that point, we use the proposed technique that gives the assessment of tweet. From the sentiment analysis, the client will understand the feedback of the services before creating a buying deal. The rest of the paper is organized as follows. Section II provides related work for dissecting previous work. In Section III, architectural overview of the proposed work and proposed algorithmic rule are briefly described. In the section IV, sentiment analysis of a product is briefly described. In section V, the simulated results and performance evaluation of our proposed method are presented. Section VI concludes the paper.

II. Related Work

In [6], the authors proposed a machine learning algorithm with existing twitter dataset to sentiment analysis. The concept of sentiment analysis using a proposed approach that automatically classified the Tweets as positive, negative or neutral. They are also using Part Of Speech(POS)-Specific polarity features and a tree kernel. In [7], a model for sentiment analysis is proposed which is based on ontology. The authors proposed a model work which intends to mix domain ontology with natural language process techniques to spot the sentiment behind judgments going to offer a description for such polarization. The methodology tests were developed by using two individual areas, digital camera, and movies [8]. In [9]-[10], the methods of sentiment analysis are trained for detection sentiment polarity which can automatically track out sentiments from different documents, blog, sentences or words. In [11], the authors developed a new method for semantic knowledge extraction from research documents and article using an integration of semantic technology, NLP, and information extraction. In [12], the authors proposed a novel method which extracts structured data from emails by using data cleaning, data extraction, and data consolidation. In [13], the authors proposed that the supervised learning approach is based on label datasets which are trained to provide meaningful outputs. To supervise the learning approach, apply Naive Bayes algorithm, maximum entropy and support vector machine which helps to achieve great success in sentiment analysis. In [14], the authors showed that they have got a maximum of 82.1% accuracy by using Naive Bayes Algorithm. In [15], the authors used K-Nearest Neighbor classifier to sentiment analysis and have got 74.74% accuracy. In [16],

978-1-7281-3060-6/19/$31.00 ©2019 IEEE

the authors were presenting a survey on sentiment analysis of Twitter data with using different techniques. They used different machine learning algorithms, such as Naive Bayes, Maximum Entropy and Vector Machine support, to sentiment analysis and demonstrate the accuracy of different sizes of features.

III. Proposed Work

This section represents our proposed work and focuses on a strategy to sentiment analysis on Twitter data. The architectural overview describing an overall process design of sentiment analysis which is shown in Fig. 1. The developed method is based on three important parts that are Data Extraction from a particular project or product, pre-processed the extracted Tweet using Natural language toolkit (NLTK) and Classifier model that calculates the sentiment of each Tweet.

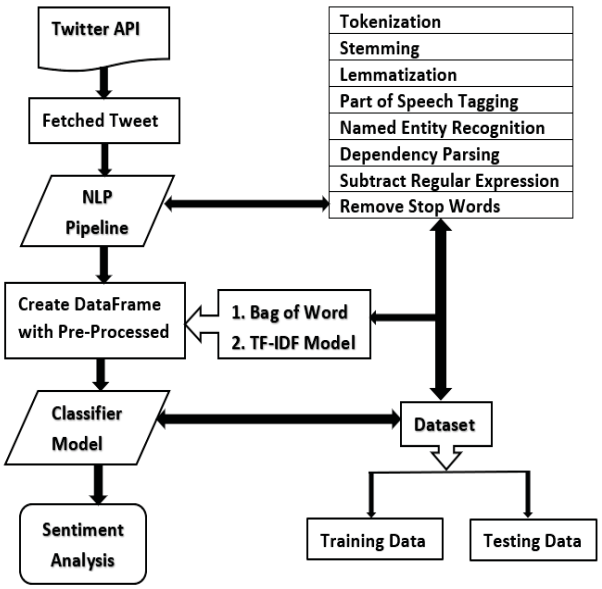


Fig. 1: Architectural overview of proposed work

To sentiment analysis, the Twitter data is using data extraction, tokenization, stemming, lemmatization, stopwords removal, parts of speech tagging, named entity recognition, create a data frame, text modeling [17], a classifier model and each of these processes has its own algorithm and packages to be processed. TF-IDF model has used to find out important words from tweets to predict sentiment. Pickle module helps to build a classifier model where pickle has done the object serialization process. To uncover the sentiments, at first we have built a classifier model by using a dataset [18] which is shown in Algorithm 1. We also have developed a natural language processing (NLP) based pre-processed data framework to filter tweets and sentiment analysis which is shown in 2. Basically, we have used BoW model to document classification and text modeling and feed this model in our method to analyze tweets. Since, BoW gives all words to have the same importance and it does not preserve any semantic information. We incorporate TF-IDF model that overrides BoW where the TF-IDF model values help to find out the most important word in tweets. And it helps to sentiment analysis with high precision. Based on the classifier model, our proposed system evaluates whether the tweet is positive or negative.

Algorithm 1 Pseudocode To build classifier Model

Input: A Dataset
Output: Create TFIDF model and Classifer Model
Method

```
1. Importing dataset
2. Serialization and de-serialization
3. save to a file and then load again
4. FOR each data in dataset :
5.     extract from and append required data into dataset
6. Create a simple binary bag of words model
7. Create a simple TF-IDF model
8. Create training and testing dataset :
9. Create classifier model by using logistic regression
10. Pickle the classifier
11. vectorizer ← TFIDFVectorizer ()
12. Unpickle the classifier and vectorizer for read operation
```

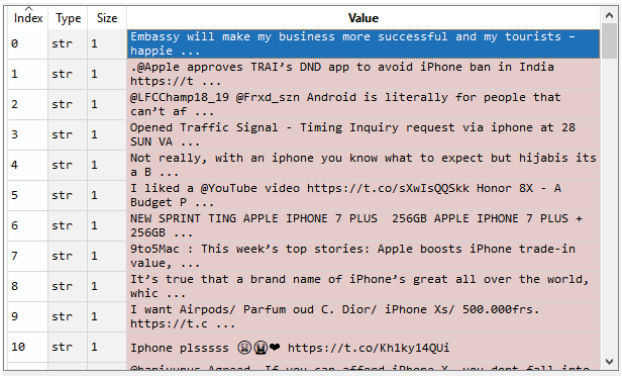
Algorithm 2 Pseudocode for Sentiment Analysis with NLP Technique on Twitter Data

Input: Fetching Tweets From Twitter
Output: Given Sentiment Polarity of Each Tweet
Method

```
1. function twitterAPI_setup():
2.     Authenticating and accessing twitter keys
3.     set argument 'iPhone'
4.     return api
5. end function
6. list_tweets = []
7. Create a dataframe list_tweets by using pandas dataframe
8. Loading TF-IDF model and classifier
9.         retrieve data into matrix form TF-IDF model
10.        load TF-IDF model
11.        load classifier
12. Importing re library
13. Initialize p ← 0 , n ← 0 ;
14. FOR tweet in list_tweets :
15.        extract required tweet from list_tweet
16. Using vectorizer and classifier model to predict sentiment
17. Print tweet with sentiment
18. IF sentiment[0] = = 1 :
19.        increase positive tweet
20. ELSE :
21.        increase negative tweet
22. Total_positive ← p
23. Total_negative ← n
```

IV. Sentiment Analysis

Our developed method for sentiment analysis from twitter on twitter data is presented in this research. As an example, we show the comparison between two the most popular device "iPhone" and "Samsung". We fetched 100 tweets on iPhone from Twitter which is shown in Fig. 2. By using NLP pipeline we have cleaned redundant information and stop words are removed from tweets to sentiment analysis which has already discussed in the previous section. To predict sentiment now we used "vectorizer" and "classifier" which is "clf" object which is created from the "tfidfmodel.pickle" and "classifier.pickle". Then we show the prediction of sentiment according to each tweet for iPhone device which is shown in Fig. 3. Now we



Index	Type	Size	Value
0	str	1	Embassy will make my business more successful and my tourists - happie ...
1	str	1	.@Apple approves TRAI's DND app to avoid iPhone ban in India https://t ...
2	str	1	@LFCChamp18_19 @Frxd_szn Android is literally for people that can't af ...
3	str	1	Opened Traffic Signal - Timing Inquiry request via iphone at 28 SUN VA ...
4	str	1	Not really, with an iphone you know what to expect but hijabis its a B ...
5	str	1	I liked a @YouTube video https://t.co/sXwIsQQSkk Honor 8X - A Budget P ...
6	str	1	NEW SPRINT TING APPLE IPHONE 7 PLUS 256GB APPLE IPHONE 7 PLUS + 256GB ...
7	str	1	9to5Mac : This week's top stories: Apple boosts iPhone trade-in value, ...
8	str	1	It's true that a brand name of iPhone's great all over the world, whic ...
9	str	1	I want Airpods/ Parfum oud C. Dior/ iPhone Xs/ 500.000frs. https://t.c ...
10	str	1	Iphone plsssss 😩😭❤ https://t.co/Kh1ky14QUi

Fig. 2: Sample fetched tweet

take a sample of a tweet in order to check our classifier model whether does it give the correct sentiment among a tweet which is shown in Fig. 4. In this tweet, an emoticon is used since we do not analysis the tweets according to emoticon so the NLP pipeline clean this emoticon and many redundant words in this tweet such as "https://t.co/c1yNjARvif" which is converted into "ynjarvif" by using Python re package for preprocessing the tweet and this word doesn't mean anything so TF-IDF model would give zero. Here the classifier model predicts the sentence as a negative sentence by giving 'zero' polarity which is shown in Fig. 5. By educing the sentence we may see that the user want to say about another device or object which is better than iPhone. So, the tweet actually expresses a negative statement about iPhone product. Therefore, our proposed technique can also give a sentiment for comparative sentences.

```
-embassy will make my business more successful and my tourists happier iwsdt nn gameinsight
iphonegames iphone : [1]
- apple approves trai dnd app to avoid iphone ban in india utbv ktwvs : [1]
- lfcchamp _ frxd_szn android is literally for people that can afford an iphone : [1]
-opened traffic signal timing inquiry request via iphone at sun valley bv se iupsip cmk guarantee
envhh mvhf : [1]
-not really with an iphone you know what to expect but hijabis its big surprise under all that fabric
wunqo lf : [0]
- liked youtube video sxwisqqskk honor a budget phone with iphone xs max bezels  : [0]
-new sprint ting apple iphone plus gb apple iphone plus gb jet black uhtbxd uopcnwugxk : [1]
- to mac this week top stories apple boosts iphone trade in value apple music on echo more mf hyr  :
[1]
-it true that brand name of iphone great all over the world which isn of urse limited to japan but do
qoondkrsfi : [1]
- want airpods parfum oud dior iphone xs frs fmkqtnbrdl : [0]
-iphone plsssss kh ky qui : [1]
```

Fig. 3: Pre-processed tweet with sentiment

Much better than an iphone😍 https://t.co/c1yNjARvif

Fig. 4: Example of a messy tweet

-much better than an iphone ynjarvif : [0]

Fig. 5: Pre-processed tweet

The ratio of the positive tweet and negative tweet according to fetched 100 tweets is given below by plotting a Bar Diagram by using Python Matplotlib package which is showing in Fig. 6.

V. Simulation and Performance Evaluation

We fetched 1000 tweet for iPhone and Samsung then calculate the total positive and negative sentiment for both devices which is shown in TABLE I. Here, we have seen that the iPhone phones are more popular than Samsung Phone. For 100 tweet for iPhone and Samsung, the comparison which is shown in TABLE II.

Here, we can see that the Samsung phones are more popular than iPhone. But The result doesn't create any major infliction,

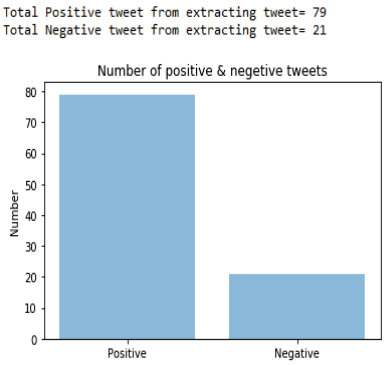


Fig. 6: Bar chart of sentiment

TABLE I: Sentiment evaluation for 1000 tweet

Output	Positive	Negative
Total Tweet for iPhone	726	274
Total Positive Tweet for Samsung	712	282

TABLE II: Sentiment evaluation for 100 tweet

Output	Positive	Negative
Total Tweet for iPhone	66	34
Total Positive Tweet for Samsung	70	30

It just gives an inference about the sentiment analysis. To get better understand and visualization, we have shown different sentiment according to a number of collecting tweet such as 50, 100, 500, 1000 respectively which is shown in Fig. 7. Two companies have launched many phones within a year where some phones have taken more popularity than other company's phone and it varies phone to phone. So, it is difficult to say one company is more popular than other company because the ratio of positive and negative tweet among two phones isn't significant. But we can say both two phone company are more popular than other phone company, as shown in Fig. 8.

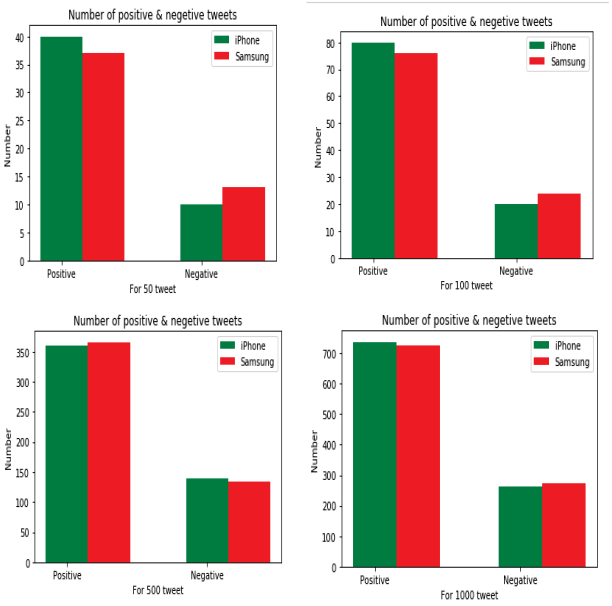


Fig. 7: Comparison between two device by sentiment

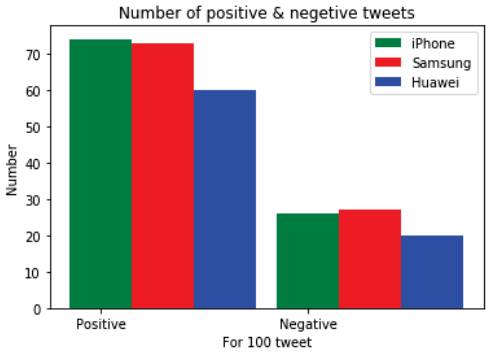


Fig. 8: Comparison between three device

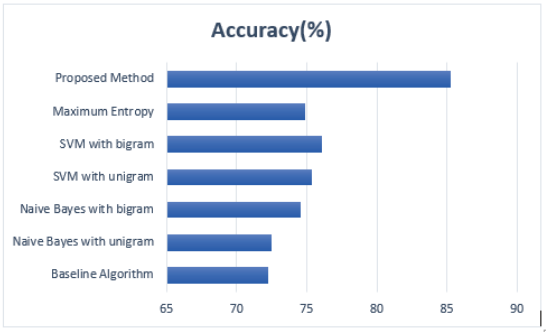


Fig. 9: Comparison of Performance Evaluation

A. *Performance Evaluation*

From TABLE III, IV, V, we have got maximum accuracy of 85.25% according to max_features, minimum document frequency (min_df), maximum document frequency (max_df). We compare our proposed method with support vector machine (SVM) [16], maximum entropy [16], naive bayes algorithm [14] and k-nearest neighbor classifier [15]. Therefore, we have taken the accuracy of 2000 features for each technique and evaluated the performance of the techniques which are more efficient for the sentiment analysis which is shown in Fig. 9. From Fig. 9, we see that our proposed technique achieved 85.25% accuracy which outperforms other techniques.

TABLE III: Accuracy for min_df=1

Max_features	min_df	max_df	Accuracy(%)
2000	1	0.1	78
2000	1	0.2	79
2000	1	0.3	81.5
2000	1	0.4	82.75
2000	1	0.5	84.75
2000	1	0.6	84.75
2000	1	0.7	84.75
2000	1	0.8	85.25
2000	1	0.9	84.75
2000	1	1.0	84.25

TABLE IV: Accuracy for min_df=2

Max_features	min_df	max_df	Accuracy(%)
2000	2	0.1	78
2000	2	0.2	79
2000	2	0.3	80.75
2000	2	0.4	82.5
2000	2	0.5	84.75
2000	2	0.6	84.5
2000	2	0.7	84.75
2000	2	0.8	85.25
2000	2	0.9	85
2000	2	1.0	84.25

TABLE V: Accuracy for min_df=3

Max_features	min_df	max_df	Accuracy(%)
2000	3	0.1	78.25
2000	3	0.2	79.25
2000	3	0.3	81.25
2000	3	0.4	82.75
2000	3	0.5	84.75
2000	3	0.6	84.75
2000	3	0.7	85
2000	3	0.8	85
2000	3	0.9	85
2000	3	1.0	84.25

VI. Conclusion

In this paper, we have developed a model to sentiment analysis which allows the processing of Twitter API streaming feed in real time and to classify its polarity to provide valuable insight in industry and users [17]. Our built classifier can be utilized as data analysis tools in NLTK. Therefore, in general we can use our proposal technique to sentiment analysis for any device, public figure or sports team that is better than any other existing model with high accuracy performance.

References

[1] I. Ahmad, "how much data is generated every minute info," *Social Media Today*, Jun 2018.
[2] O. Alonso, K. Shiells, H. J. Lee, and C. Carson, "System and method for generating social summaries," Jan 2019.
[3] K. C. Yang and Y. Kang, "Microblogs, jasmine revolution, and civil unrest: Reassessing the emergence of public sphere and civil society in people's republic of china," in *Censorship, Surveillance, and Privacy: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2019, pp. 1153–1178.
[4] Z. Lv, H. Song, J. Lloret, D. Kim, and J.-N. De Souza, "Ieee access special section editorial: Big data analytics in the internet-of-things and cyber-physical systems," *IEEE Access*, vol. 7, pp. 18 070–18 075, 2019.
[5] N. Book, "Natural language toolkit," *NLTK 3.4 documentation*. [Online]. Available: www.nltk.org/
[6] V. Sahayak, V. Shete, and A. Pathan, "Sentiment analysis on twitter data," *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, vol. 2, no. 1, pp. 178–183, 2015.
[7] F. Ceci, A. L. Goncalves, and R. Weber, "A model for sentiment analysis based on ontology and cases," *IEEE Latin America Transactions*, vol. 14, no. 11, pp. 4560–4566, 2016.
[8] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.
[9] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of web services," *Information Sciences*, vol. 311, pp. 18–38, 2015.
[10] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.
[11] R. Upadhyay and A. Fujii, "Semantic knowledge extraction from research documents," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016, pp. 439–445.
[12] S. S. A. Q. Mahlawi, "Structured data extraction from emails," *2017 International Conference on Networks Advances in Computational Technologies (NetACT)*, vol. 37, pp. 323–328, 2017.
[13] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 502–518.
[14] N. P. M. Vadivukarassi and P. Aruna, "Sentimental analysis of tweets using naive bayes algorithm," *World Applied Sciences Journal*, vol. 35, no. 1, pp. 54–59, 2017.
[15] S. Pal and S. Ghosh, "Sentiment analysis using averaged histogram," *International Journal of Computer Applications*, vol. 162, no. 12, 2017.
[16] V. Kharde, P. Sonawane *et al.*, "Sentiment analysis of twitter data: a survey of techniques," 2016.
[17] R. Hasan, "Sentiment analysis with nlp on twitter data," Jun 2019. [Online]. Available: https://github.com/Rakib-Hasan031/Sentiment-Analysis-with-NLP-on-Twitter-Data
[18] L. L. Bo Pang, "Software and data." *Cornell NLP*. [Online]. Available: nlp.cornell.edu/data/