QUALITY ANALYSIS OF TOP COLLEGES USING THE TWITTER DATA

#Dr.M.Lilly Florence M.Tech., Ph.D.,*R.Uma Maheswari*G.Vaishnavi

#Professor*Final Year B.E(CSE)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ADHIYAMAAN COLLEGE OF ENGINEERING

HOSUR, TAMIL NADU, INDIA

***Abstract*—** In today's world, opinions and reviews play a critical factor in formulating our views and influencing the success of the product or service. With the extreme growth of social media, people often express their views on one of the popular social media, named twitter. Twitter presents a challenge for analysis because of its humongous and disorganized nature. The work in this report Quality Analysis of Top Colleges Using Twitter Data is to find the top colleges in the country, considering the tweets from one of the popular social media named twitter. In the process of analysis, many preprocessing techniques can be applied to the data that is generated by twitter. The analysis is done using the python programming, applying the TextBlob for preprocessing and Naïve Bayes for classification the result will be shown in the form of a graph. Based on the percentage of the positive tweets the ranking of the colleges is done

Key Words—Quality Analysis, TextBlob, Naïve Bayes, Positive reviews, Ranking of Colleges

I. INTRODUCTION

With the huge amount of increase in web technologies, the no of people expressing their views and the opinion via the web are increasing. This information is useful for everyone like businesses, governments and, individuals with 500+ million tweets per day, twitter is becoming a major source of information. The raw tweets are given as input. We automate the process of tweet extraction and categorizing it into three categories i.e. positive, negative and neutral tweets. The content in twitter generated by the user is about different kinds of products, events, people and, political affairs[1].

Performing quality analysis on tweets is considered best due to the following reasons:

- Analysis of real-time can be done.
- A huge variety for performing the analysis [2].

Quality analysis can be defined as a process that extracts the attitudes, opinions, views, and emotions from the text, speech, tweets, and database source through NLP it is also known as opinion mining, it detects whether a piece of writing is positive, negative or neutral[3]. Today most people use social networking sites to express their opinion about something. Companies have been receiving polls about the products they manufacture. The quality analysis is done using various machine learning techniques. Analysis can be done at the document, phrase and sentence level. In the document level, the entire document is taken then it is analyzed whether the content is positive, negative or neutral. In phrase level, the analysis of phrases in a sentence is taken into account to check the polarity. In sentence level, each sentence is classified into the number of classes[4]. The goal of quality analysis is to harness this data to obtain important information regarding public opinion, which would help make smarter business decisions, political campaigns, and better product consumption [5]. Here we are doing the sentence-level analysis, all the sentences are taken into the .csv file then pre-processing apply on those sentences and using a machine learning algorithm that data is classified[6].

In today's world, roughly 34,582,000 out of an estimated 176,805,000 of the 18-23-year-old age group in India receive higher education which equates to about 19.56% of the age group. Much reputed government and private colleges in India aim towards providing class education to their students and follow different ideologies, pedagogies and examination procedures. It is highly useful for the interested student to evaluate the choices available to him/her in selecting a college that not only furnishes the student with the desired academic or professional prowess but also equips him/her with the right kind of learning tools according to his/her capabilities. The tweets about the colleges have been collected and analyzed to find the user's opinion on the perception of these colleges, the ranking of the college is found[7].

II. RELATED WORKS

Qi Gu, Eugene Santos Jr, Eunice E. Santos has done modeling opinion dynamics in the social network. Opinion dynamics is a complex procedure that entails a cognitive process when dealing with how a person integrates

influential opinions to form a revised opinion. In this work, we present a new approach to model opinion dynamics by treating the opinion on an issue as a product inferred from one's knowledge bases, where the knowledge bases keep growing and updating through social interaction. A general impact metric is proposed to evaluate the likelihood of a person adopting the opinions of others. Specifically, a set of domain-independent influential factors is selected based on social and communication theories, but the weights of these factors are missing. Though the opinions from different actors are not integrated linearly like traditional methods, we show that the factor weights can be efficiently learned via regression. We validated the effectiveness of our model by comparing it against a baseline model on both synthetic and real datasets[8].

Arora, Li, and Neville used Lexicon based Sentiment analysis on various smartphone brands to judge their popularity and reviews in the range of sentiment scores from -6 to 6[9].

Twitter has received much attention recently. An important characteristic of Twitter is its real-time nature. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo investigated the real-time interaction of events such as earthquakes in Twitter and propose an algorithm to monitor tweets and to detect a target event. To detect a target event, the authors devised a classifier of tweets based on features such as the keywords in a tweet, the number of words, and their context. Subsequently, the authors produced a probabilistic spatiotemporal model for the target event that can find the center of the event location. They regarded each Twitter user as a sensor and apply particle filtering, which is widely used for location estimation. The particle filter works better than other comparable methods for estimating the locations of target events. As an application, the authors developed an earthquake reporting system for use in Japan.

Because of the numerous earthquakes and a large number of Twitter users throughout the country, the author can detect an earthquake with high probability (93 percent of earthquakes of Japan Meteorological Agency (JMA) seismic intensity scale 3 or more are detected) merely by monitoring tweets. The systems detect earthquakes promptly and notification is delivered much faster than JMA broadcast announcements[10].

Neethu M. S. and Rajasree R used twitter posts on electronic products, compared the accuracy between different machine learning algorithms and further improved the accuracy by replacing repeated characters with two occurrences, including a slang dictionary and considering emoticons[11].

Researchers have also been working upon the prediction of the accuracy of the tested dataset using Machine Learning algorithms. Kanakaraj and Guddeti used Natural Language Processing Techniques for sentiment analysis and compared Machine Learning Methods and Ensemble Methods to improve on the accuracy of the classification[12].

Also, the area of neural networks has been investigated for performing sentiment analysis on benchmark datasets consisting of online product reviews. Bespalov, Bai, Qi, and Shokoufandeh carried out binary classification on Amazon and TripAdvisor datasets using a Perceptron classifier and obtained one of the lowest error rates among their experiments of 7.59 and 7.37 on the two datasets respectively[13].

III. SYSTEM ANALYSIS

A. Existing System

In Existing Method, three of the premier colleges in India, namely the All India Institute of Medical Sciences (A.I.I.M.S.), the Indian Institute of Technology (I.I.T.) and the National Institute of Technology (N.I.T.) have been analyzed to find the user's opinion on the perception of these colleges and the magnitude of these opinions[14]. The authors used different machine learning algorithms like Support Vector Machine and Artificial Neutral Networks and Naïve Bayes to measure accuracy[15].

B. Proposed System

In the proposed system the tweets about the various college are collected from twitter. The quality analysis on the tweets is done to find the percentage of positive, negative and neutral tweets. Based on the percentage of positive tweets the ranking of the college is found. . Based on the percentage of the positive tweets the ranking of the colleges is done[16].

IV. SYSTEM ARCHITECTURE

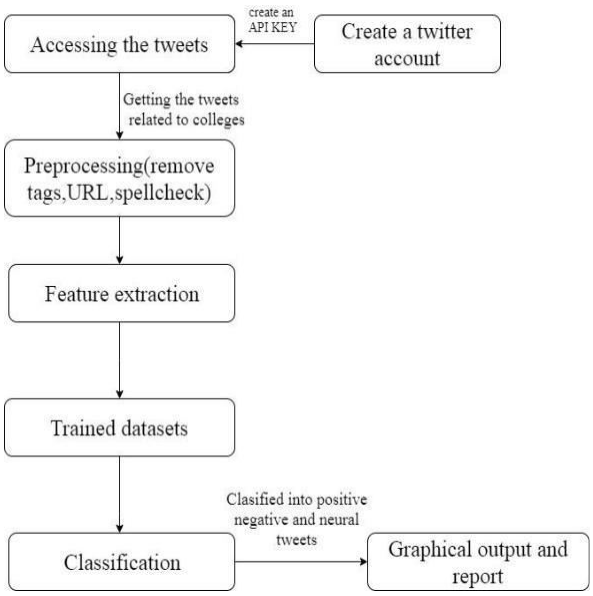

Figure 1: Architecture design

Firstly, we want to create a twitter account to get the API, secret and the consumer key to access the tweets. Then we will collect the datasets of the colleges. The preprocessing is done to remove the tags, URL, and spellcheck. After preprocessing the tweets are classified into positive,

negative and neutral tweets and the result is shown in the form of a graph. The ranking is done depending upon the percentage of the positive tweets i.e the college which has the highest positive tweets is ranked as the top college. Based on the percentage of the positive tweets the ranking of the colleges is done[17].

A) Getting the API

The first step is to create a twitter account, for accessing the tweets you need four keys

- Consumer Key
- Consumer Secret Key
- Access Token
- Access Token Key[18].

For generating the above keys follow the steps given below.

- Go to the https://apps.twitter.com/app/new and login if necessary.
- Enter your application name, description and your website address making sure to enter the full address including the https://. You can leave the call back URL empty.
- Accept the conditions and submit the form by clicking the create your twitter application.
- After creating your application click on the tab that says keys and access tokens then you have to give access to your account to use this application. To do this select the create my access token button.
- Lastly copy all the keys i.e consumer key(API key), consumer secret access token and access token secret from the screen into your plugins twitter options page and test[19].

After generating the keys the datasets are collected using the python code[20].

B)Preprocessing

Every second, on average, around **6,000 tweets** are tweeted on Twitter, which corresponds to over **350,000 tweets** sent per minute, **500 million tweets** per day and around **200 billion tweets** per year[21]. Unlike other text documents, Twitter may be a domain where people use their freedom to express messages or comments flexibly. Twitter acquired a lot of raw information that may or may not find a use for the application. Several attributes have been identified for a Twitter 53 status update or tweets. The maximum length of the Twitter message is 140 characters which may include user-mention, hashtag, URL, etc. The frequency of elongated words, slangs, acronyms, and emoticons is much higher than any other domain[22].

Pre-processing is fundamental to all Natural Language Processing (NLP) tasks. Steps needed for pre-processing of text, in general, depending on the targeted requirement or application. Here tweets are pre-processed for remove unwanted characters such as URL, hashtag, etc.

Here preprocessing is done using the TexttBlob. TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, feeding the keywords, translation, and more. Below are some of the features of the TextBlob[23].

Noun phrase extraction
Part-of-speech tagging
Language translation and detection powered by google translate
Tokenization(splitting text into words and sentence)
Word and phrase frequencies Parsing
Wordinflection and lemmatization Spelling correction
Wordnet integration

C) Classification

After preprocessing, all the tweet terms containing the keyword are classified into positive, negative and neutral tweets. Naïve Bayes algorithm is used for classification. Here we are having good words and bad words datasets. By comparing with this, we can classify the tweets into positive, negative and neutral tweets. The percentage of the positive, negative and neutral tweets is found. Depending upon the percentage of the positive tweets the ranking of the college is done.

Naïve Bayes classification is nothing but applying the Bayes rule for forming the classification probabilities[24].

For a document d and class c

$$P(c/d)=P(d/c)*P(c)/P(d)$$

P(c/d) is read as the probability of a class c given a document d, it is also called as the posterior probability. The P(d/c) is called the likelihood, it is the probability of document d given a class c. P(c) is the prior, it is the original label of the document being positive, negative or neutral. P(d) is the normalization constant, it is used to ensure that the outcome can be represented in the probability distribution[25].

D) Graphs and results

A graphical presentation is a diagram or graph that represents a set of data. The graphical representation is the visual display of data which will help us to present data in a meaningful way and it provides data that is very easy to understand and helps management to make decisions[26]. The graph is generated between positive, negative and neutral tweets. It is based upon the count of the positive, negative and neutral tweets[27].

Based on the percentage of the positive tweets the ranking of the colleges is done[28].

V. IMPLEMENTATION

In this section, we consider the quality analysis of top colleges using the twitter data. Here using the API the twitter data are accessed and the datasets related to the colleges are collected[29]. Then using the preprocessing the noisy data are removed[30]. Using the classification the tweets are classified into positive, negative and neutral tweets. Depending upon the percentage of the tweets the ranking of the college is done. The college which has the highest positive percentage is ranked as the top college[31]. The result is shown in the form of a graph for easy understanding.[32].

VI. RESULT

	A	B	C	D	E	F	G	H
1	Tweet ID	Tweet Dat	Screen Na	user Nam	Tweet Te	Location	Retweet Count	
2								
3	1.21E+18	2020-01-0	lindseyfai	lindseyfai	Howabusi	Kingston,(	1	
4								
5	1.21E+18	2020-01-0	BALIGoLar	BALIGoLar	Afantastic	England,U	0	
6								
7	1.21E+18	2020-01-0	btivgdc	conwuezy	RT@thestevefund:ð		13	
8								
9	1.21E+18	2020-01-0	Anirudh_4	AnirudhBI	RT@go4a	Delhi	13	
10								
11	1.21E+18	2020-01-0	Commerc	Philadelp	RTACLUOurreproduc		0	
12								
13	1.21E+18	2020-01-0	latestly	LatestLY	BharatBar	Mumbai,I	0	
14								

Figure 2: Datasets of the colleges

	Tweets	len	College id	Date	Source	Likes	RTs	SA
0	Interested in exploring the mystery of how hum...	140	1015226873855598592	2018-07-06 13:32:03	Buffer	0	0	1
1	Being able to communicate well in English allo...	139	1014961376949886976	2018-07-05 19:57:03	Buffer	0	0	1
2	You have drive. You have focus & determina...	133	1014857206750162944	2018-07-05 13:03:07	Buffer	0	0	1
3	Do you have a dream of working and living abro...	140	1014544123578200066	2018-07-04 16:19:03	Buffer	0	0	0
4	RT @IUSONIndy: You can now complete the BSN ac...	140	1014296488236090881	2018-07-03 23:55:02	Buffer	0	6	1
5	Believe it or not, every career field could be...	139	1014179979163328512	2018-07-03 16:12:04	Buffer	0	0	1
6	RT @KatieEnvin: You didn't miss a chance to co...	140	1013934092184940545	2018-07-02 23:55:00	Buffer	0	1	1
7	Want to accelerate your degree even more? Buil...	140	1013860623741018113	2018-07-02 19:03:03	Buffer	0	0	1
8	RT @RCNJadmissions: Learn more about earning y...	140	1013468030893600512	2018-07-01 17:03:02	Buffer	0	2	1
9	RT @WMUGradCollege: Congratulations, Joshua Wh...	140	1013135835242299392	2018-06-30 19:03:00	Buffer	0	3	0

Figure 3: classified datasets

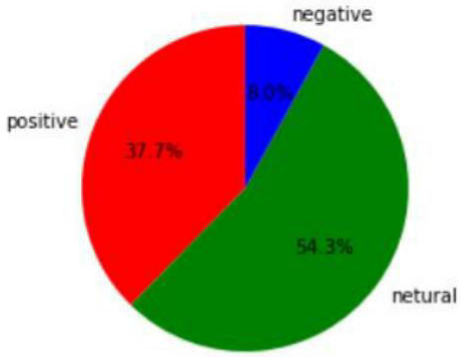

Figure 4:Graphical representation of the percentage of Tweets

	University
0	Indian Institute of Technology Bombay
1	Indian Institute of Technology Madras
2	Indian Veterinary Research Institute
3	Indian Institute of Science
4	Sri Sri University
5	Narsee Monjee Institute of Management and High...
6	Flame University
7	Indian Institute of Technology Kanpur
8	ISBM University
9	The Neotia University

Figure 5: Ranking of the colleges based on the tweets

VII. CONCLUSIONS

In conclusion, Indian Institute of Technology,Bombay is the most positively talked about college among the premier institutes of India on twitter[33]. Since we are using the tweets for finding the ranking of the colleges, we have used the Naïve Bayes algorithm for the classification, because it is used to access a large number of datasets and provides the result with high accuracy[34].

Quality analysis is an effective way of classifying the opinions given by the people related to any topic, service or product. Automation of this task makes it easier to deal with the huge amount of data being generated by the social website like in twitter real-time analysis[35]. The Naïve Bayes provides high accuracy because each time a decision with the highest probability is made[36].

VIII. REFERENCES

[1] Nehal Mamgain, Ekta Mehta, Ankush Mittal, Gaurav Bhatt, Sentiment analysis of top colleges using the twitter data, International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT),2016.

[2] Arora D., Li K.F., and Neville S.W., Consumers' sentiment analysis of popular phone brands and operating system preference using Twitter data: A feasibility study, 29th IEEE International Conference on Advanced Information Networking and Applications, pp. 680-686, Gwangju, South Korea, March 2015

[3] Choi C., Lee J., Park G., Na J. and Cho W., Voice of customer analysis for internet shopping malls, International Journal of Smart Home: IJSH, vol. 7, no. 5, pp. 291-304, September 2013

[4] Kanakaraj M., Guddeti R M.R., Performance Analysis of Ensemble Methods on Twitter Sentiment Analysis using NLP Techniques, 9th IEEE International Conference on Semantic Computing, pp. 169-170, Anaheim, California, 2015

[5] Bahrainian S.-A., Dengel A., Sentiment Analysis and Summarization of Twitter Data", 16th IEEE International Conference on Computational Science and Engineering, pp. 227-234, Sydney, Australia, December 2013

[6] Pak A. and Paroubek P., Twitter as a Corpus for Sentiment Analysis and Opinion Mining, 7th International Conference on Language Resources and Evaluation, pp. 1320-1326, Valletta, Malta, May 2010

[7] Shahheidari S., Dong H., Bin Daud M.N.R., Twitter sentiment mining: A multidomain analysis, 7th IEEE International Conference on Complex, Intelligent, and Software Intensive Systems, pp.144-149, Taichung, Taiwan, July 2013

[8] Neethu M. S. and Rajasree R., Sentiment Analysis in Twitter using Machine Learning Techniques, 4th IEEE International Conference on Computing, Communications and Networking Technologies, pp. 1-5, Tiruchengode, India, 2013

[9] Bespalov D., Bai B., Qi Y., and Shokoufandeh A., Sentiment classification based on supervised latent n-gram analysis, 20th ACM international conference on Information and knowledge management, pp. 375-382, New York, USA, 2011

[10] Jotheeswaran J. and Koteeswaran S., Decision Tree-Based Feature Selection and Multilayer Perceptron for Sentiment Analysis, Journal of Engineering and Applied Sciences, vol. 10, issue 14, pp. 5883-5894, January 2015

[11] Socher R., et al, Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, October 2013.

[12] dos Santos C. N. and Gatti M., Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts, 25th International Conference on Computational Linguistics, pp. 69-78, Dublin, Ireland, August 2014.

[13] Segaran T. and Hammerbacher J., Beautiful Data: The Stories Behind Elegant Data Solutions, Beijing: O'Reilly, 2009

[14] Nielsen F.A., Making sense of microposts, Finn ÅrupNielsenblog,https://finnaarupnielsen.wordpress.com/tag/sentimentanalysis/

[15] Koto F. and Adriani M., A Comparative Study on Twitter Sentiment Analysis: Which Features are Good?, Natural Language Processing and Information Systems, Lecture Notes in Computer Science vol. 9103, pp. 453-457, June 2015