International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Extraction and Analyze Text in Twitter using Naive Bayes Technique.

Ameen Abdullah Qaid Aqlan, B.Manjula

Abstract—there are several topics and areas that are at an advanced stage of interest and research around the world because of their importance and usefulness to humanity, including the sentiment analysis. By studding of sentiment analysis (SA), one can learn about the mysterious things and different feelings of others. The purpose of all of this is to know the pros and cons about a product or anything else and correct the negatives in future that are found. In our research, we have benefited from social media sites, especially Twitter, in collecting data about the iPhone 11 product to see how satisfied customers are about this product. We collected a lot of different opinions using API and then transferred them to an information bank. In our research we used the famous Naive Bayes (NB) algorithm and had an active role in classifying reviews and sorting them and knowing the pros and cons, where we got good results compared to previous works which are as follows: precision 80, recall 83, f1 score 81, accuracy 80.25.

Keywords—Sentiment Analysis, Data Mining, Classification Reviews, NB Algorithm, Twitter.

I. INTRODUCTION

In light of the continuous escalation and rapid increase in data, it is imperative that we find modern ways to research and analyze the opinions of people and commentators and evaluate any tweet, opinion, purpose or commodity such as (foundation, company, university, organization, etc.) using classification technology and automated science, which is a major task For the process of exploring the data used in the broader process is the exploration of knowledge. We note that between 2004 and 2019 there was a huge revolution in data mining on social media sites, with many countries adopting a special budget for research and development in data analysis [1]. According to a previous study we conducted many study researches in the field of data analysis we found an accelerated increase and a development revolution in this area.

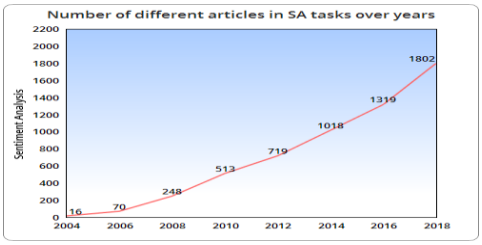

Figure 1 Number of different articles in SA tasks over years

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Ameen Abdullah Qaid Aqlan*, Department of Computer Science, Kakatiya University, Warangal, Telangana, India.
B.Manjula, Department of Computer Science, Kakatiya University, Warangal, Telangana, India.

Wars and counter-terrorism has been a driving force for data analysis and get satisfactory results [2]. The research studies also focused on people's responses and comments about products displayed on the Internet and prophecies in global and financial markets [3, 4]. This search helps the user to access the evaluation of other users through their tweets and feedback on the internet for getting immediately and automatically opinion and then process of analysis opinions using the appropriate algorithms for this purpose. There are several classifications in data exploration those are:

1-Classification
2-Prediction
3-Summarization
4-Clustering
5-Association
6-Anomaly Detection

Our research uses the classification feature that is classified within the field of forward-looking (predictive learning) learning based on previous information in order to predict future information.

The classification process is useful in extracting or developing models that define categories or rows for different purposes depending on the characteristics of these purposes, which have been learned in advance. There are many algorithms used in data exploration and the appropriate algorithm is chosen depending on the nature of the problem, in the following will mention the most important algorithms used.

1- Nearest Neighbor
2- Decision Trees
3- Naïve Bayesian
4- Clustering
5- Association Rules
6- Neural Network
7- Time Series
8- Support Vector Machines

In our research, we use the Naive Bayes algorithm, which is one of the most important algorithms in the field of data exploration, characterized by the ease of application and the ability to deal with big data and this algorithm is based on the theory of the scholar Bayes.

Most human activities depend on the opinions formed by people because they have a great influence on our beliefs and behavior and our perception of reality and the choices that we make, so when you have a problem or want to make a specific decision we ask others to express their opinions, this method is not exclusive to Only people But also on organizations, institutions and others, in the past few years the ways in which people express their opinions and feelings have changed radically thanks to the emergence of social networks, virtual society and other media and social media on the Internet.

Internet usage has increased in recent years; according to statistics and Estimates Internet World Stats even Mid-Year of 2019 the number of internet users has reached about 4,536,248,808 users; we'll illustrate this data in table 1. [5].
The discovery of knowledge from a huge amount of unstructured data on the web has become an important challenge nowadays because of its importance in various areas of life, for example, it can be a major factor for marketers who want to leave behind a picture or identity in the minds of their customers for their products and brands. Automated analysis of opinions actually involves a deep understanding of the natural language text by the organs and this goal is close to verification. Analysis of opinions and all related concepts such as sentiment, assessments, attitudes and emotions are all subjects for study by the Technique of analysis [6]. With the rapid and simultaneous growth of social media and communications such as forums and social networks, for the first time in human history we have had so much information and data recorded and stored digitally, Since 2000 sentiment analysis has a development in the processing of natural languages more than any other field.

2. Proposed Methodology to Analysis Opinions

The process of sentiment analysis and opinion is the process of studying people's opinions and their behavior and expressions and the direction of certain issues using the computer, and we have three levels of analysis those are:

-Analysis at the document level.

At this level, the document is treated as single information and the goal is to classify the document positively or negatively based on the feelings and opinions in it.

- Analysis at the sentence level.

The objective of this type of analysis is to categorize the feelings expressed in a sentence, and the work is divided into two phases, the first stage is determined whether the sentence contains feelings and opinions, and in the second stage these opinions are classified negatively or positively and therefore the sentence is determined if it carries a positive or a negative opinion.

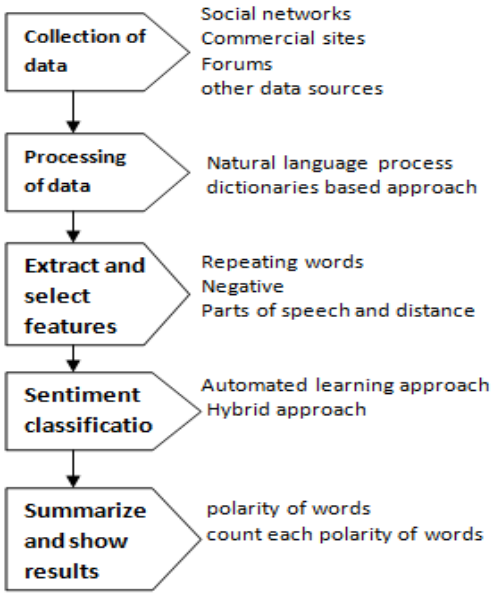

figure 2 Proposed Methodology to Analysis Opinions.

- Analysis at the level of feature or recipe

The purpose of this type of analysis is to categorize opinion or feeling, taking into account a specific description of the entity in question. The first stage of this type of analysis is the identification of the entity and its characteristics, the second stage is the analysis and classification of feelings and opinions associated with each description, the opinion holder can express the opinions and feelings in different descriptors for the same entity.

Table 1 Growing number of Internet users around the world

WORLD INTERNET USAGE AND POPULATION STATISTICS 2019 Mid-Year Estimates						
World Regions	Population (2019 Est.)	Population % of World	Internet Users 30 June 2019	Penetration Rate (% Pop.)	Growth 2000-2019	Internet World %
Africa	1,320,038,716	17.1 %	522,809,480	39.6 %	11,481 %	11.5 %
Asia	4,241,972,790	55.0 %	2,300,469,859	54.2 %	1,913 %	50.7 %
Europe	829,173,007	10.7 %	727,559,682	87.7 %	592 %	16.0 %
Latin America / Caribbean	658,345,826	8.5 %	453,702,292	68.9 %	2,411 %	10.0 %
Middle East	258,356,867	3.3 %	175,502,589	67.9 %	5,243 %	3.9 %
North America	366,496,802	4.7 %	327,568,628	89.4 %	203 %	7.2 %
Oceania / Australia	41,839,201	0.5 %	28,636,278	68.4 %	276 %	0.6 %
WORLD TOTAL	7,716,223,209	100.0 %	4,536,248,808	58.8 %	1,157 %	100.0 %

II. COLLECTION OF REVIEWS

The language of R Program is one of the important languages in the field of programming, especially in the field of data analysis and reviews, where it became Basic programming language in many universities and scientific research centers, Where became their use and reference in published articles and scientific fields increasingly exorcised, this is an issue about The fact that it is a free open source language. R Program is a suitable ground for collecting data and opinions from Twitter using APPLICATION PROGRAM INTERFACE (API) [7, 8]. API is an open app that supports developers on Twitter and allows them to upload data and comments about people, entities or products. In this paper, we have chosen the Latest release mobile from apple (iphone 11 pro).

Where he has gained international acclaim and widespread fame around the world because of his new advantages, there have been many controversies through the Twitter platform about this mobile, There are who praised it, and there are who have criticized it. We have brought more than 2,000 tweets from Twitter to analysis them and fine out the percentage of positive and negative opinions.

Figure 3. Word cloud of tweets

Pre-processing of reviews

The reviews of users in Twitter contain symbols, terms and other links, which is an obstacle when analyzed such as @, #, URL and space so we had to get rid of these impurities and symbols in order to get a useful sentence and words. Sometimes reviews may need to correct spelling mistakes in order for the dictionary to recognize them correctly. In this paper, we are interested in researching and analyzing texts written in English as it is the official language of most Twitter users.

For example

Input Tweets {RT @EMAY4K: I flip my phone over like this to let you niggas know I have an iPhone 11 Pro Max in midnight green https://t.co/QqorOXDeZE}.

Output tweets

{'I','flip','my','phone','over','like','this','let','you','niggas ','know','I','have','iphone','pro','max','midnight','green'}

In the output of text processing we got useful words that have many connotations and meanings that we can categorize into positive or negative. We have been careful to purify the text from the distances and symbols that may constitute an obstacle and an imbalance in analysis such as (hashtag @, star *, space, #, https). Some symbols are important and indicate the meaning and deletion may affect the meaning of the text, so we made sure to define these symbols in the dictionary so as not to be neglected in the processing, we will mention some of the in the following table 2.

Table 2 example of negative and positive emotion symbols

Positive	Emotion	Negative	Emoticons
:-) ,:),:D,:o),:]	happy	:-(, :(, >:[, :<	Sad
(^v^),(^u^),(^o^),^-^	happy	T-T , T^T , '_' ,	Sad
:-D,8-D,XD,=3,B^D	laugh	:'-(, :'(	Cry

3. Naïve Bayes Classifier

It is a guided statistical algorithm (based on the mathematical census method) used to predict the probability of results to solve the problems of both classification and forecasting.

Classification Naive Bayes assumes the existence of a special characteristic (target) whose existence has nothing to do with other qualities, and the other attributes are considered as random variables. Using the algorithm Naive Bayes probabilities according to the following laws.

$$P(C|A) = \frac{P(A,C)}{P(A)} \qquad (1)$$

$$P(A|C) = \frac{P(A,C)}{P(C)} \qquad (2)$$

The final formula of Naive Bayes will be as follows:

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)} \qquad (3)$$

An example that shows the data classification process by using Naive Bayes

We suggested to finding a set of positive and negative words to apply our equation and show how to Purification and classify the data into positive and negative, these words are reviewed in the following table.

Table 3 A set of positive and negative words

	Documents	Words	Class
	1	Like happy good	Positive
	2	Good like good	Positive
	3	Cool fantastic	Positive
	4	Bad good hate	Negative
Test	D5	Good like bad	?????

First step, counting both positive and negative words in the table.

$$P(\text{positive}) = \frac{DOCUMENT\ COUNT\ (POSITIVE)}{N_{document}} = \frac{3}{4}$$

$$P(\text{negative}) = \frac{DOCUMENT\ COUNT\ (NEGATIVE)}{N_{document}} = \frac{1}{4}$$

Second step Finding the result of the test text in D5

$$P(\text{like}\,|\text{positive}) = \frac{COUNT\ (like,positive)+1}{\Sigma_{W\in V}\ \text{COUNT(like,positive)}+|v|} = \frac{2+1}{8+7} = 0.2$$

$$P(\text{good}\mid\text{positive}) = \frac{COUNT\ (good,positive)+1}{\Sigma_{W\in V}\ \text{COUNT(good,positive)}+|v|} = \frac{3+1}{8+7} = 0.266$$

$$P(\text{bad}\mid\text{positive}) = \frac{COUNT\ (bad,positive)+1}{\Sigma_{W\in V}\ \text{COUNT(bad,positive)}+|v|} = \frac{0+1}{8+7} = 0.0666$$

$$P(\text{bad}\mid\text{negative}) \frac{COUNT\ (bad,negative)+1}{\Sigma_{W\in V}\ \text{COUNT}(bad,negative)+|v|} = \frac{1+1}{3+7} = 0.2$$

$$P(\text{good}\mid\text{negative}) \frac{COUNT\ (terrible,negative)+1}{\Sigma_{W\in V}\ \text{COUNT}(terrible,negative)+|v|} = \frac{1+1}{3+7} = 0.2$$

P (like| negative) $\frac{COUNT\ (hate,negative)+1}{\Sigma_{W\in V}\ \text{COUNT}(hate,negative)+|v|} = \frac{1+1}{3+7} = 0.2$

Third step Collect the result of the test row for both cases

P (positive | D5) = $\frac{3}{4}$ * 0.2*0.266*0.066 = 0.0026

P (Negative | D5) $\frac{1}{4}$ *0.2*0.2*0.2 = 0.002

It seems to us through this process that positive texts are greater than negative texts.

III. RESULT ANALYSIS

In this paper, we have been able to obtain more than 1700 tweets from Twitter for study, analysis and evaluation. A new product of apple has caught our eye, it's (iPhone 11), and it has advantages and features that were not found in previous models. But there are users who see in this product positive and some see negatives, these expressions appear in their Tweets on Twitter, We had our role in research and investigation about these tweets and collected 1700 of them from 01/11/2019 until 30/12/2019 and then saved it in the information bank in our system.

Algorithm 1: importing-libraries

begin
import all the required libraries
numpy, re, nltk, tweetpy, load-files,
stopwords, CountVectorizer, train_test_split
confusion_matrix, OAuthHandler
naive bayes, classification-report
accuracy- score
End.

We brought all the 1700 data stored in the information bank to sort it out and know the number of positive reviews and the number of negative reviews. Classifying these data to positive and negative is somewhat difficult because many users around the world sometimes give their opinions in the Slang and some other words from languages other than English. We've overcome all these constraints, and we've got a data classification as the following results.

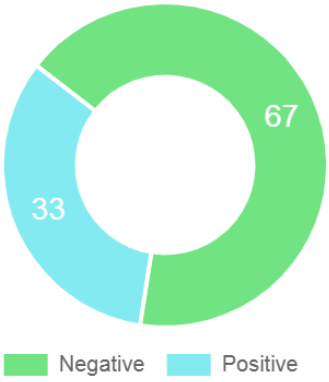

figure 4 number of positive and negative

Precision and accuracy they are a two important elements we should be more focused them when taking statement measurements. Both of them reflect of great importance as they accurately determine how close the measurement is to the actual value. But the accuracy is more accurately [10], it reflects how close the measurement is to the known value. While precision reflects how reproducible measurements are, even if they are far from the accepted value. In our classier, we've got better results compared to previous works. We've got the following results. Precision 80, recall 83, f1 score 81, accuracy 80.25.

$$\text{Precision} = \frac{true\ pos}{true\ pos+fals\ pos} \quad (4)$$

$$\text{Recall} = \frac{true\ pos}{true\ pos+fals\ neg} \quad (5)$$

$$\text{F1 score} = 2.\frac{precision.recall}{precision+recall} \quad (6)$$

In this paper we completed all important steps such as collecting data, building a dictionary and classifying data so that we could get good results, and we actually got good results compared to previous work, where we compared our work with one of the previous researches and our study was good. We will explain the comparison between the two studies in the following figure.

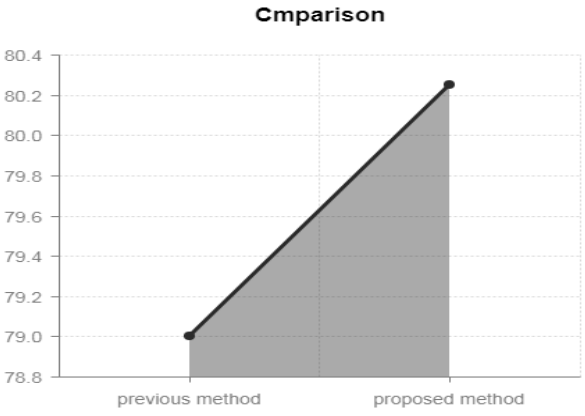

Figure 5. comparison previous work and proposed work.

IV. CONCLUSIONS

Our research coincides with a growing global interest in analyzing the big data stored on the Internet. We've collected about 1700 tweets using the API app, which allows researchers and developers to collect data for analysis and use to fix mistakes and improve the product. We have attracted our attention to the iPhone 11 product, which has a great reputation and new features that no product has ever acquired before. We used our approach of analyzing and classifying the reviews and we got good results and it was as follows: precision 80, recall 83, f1 score 81, accuracy 80.25.

REFERENCES

1. M V. Mäntylä and other, "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers", Computer Science Review, 2017.
2. Pete Burnap and other, "Tweeting the terror: modeling the social media reaction to the Woolwich terrorist attack", DOI 10.1007/s13278-014-0206-4, 2014.
3. K Dave, S Lawrence and David M. P," Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", NEC Laboratories America, Princeton, New Jersey, 2003.
4. A K Nassirtoussi and other," Expert Systems with Applications"doi.org/10.1016/j.eswa.2014.06.009, 014.
5. Internet World Stats, https://www.internetworldstats.com/stats.htm

6. H Tang, S Tan and Xueqi C," A survey on sentiment detection of reviews" doi:10.1016/j.eswa.2009.02.063, 2009
7. Pollyanna Gonçalves, "Comparing and Combining Sentiment Analysis Methods", doi.org/10.1145/2512938.2512951.,2013.
8. https://developer.twitter.com/en/premium- apis Twitter apps
9. Tom M. Mitchell, "Machine Learning" book, 1997.
10. Y Jiao and Pufeng Du, "Performance measures in evaluating machine learning based bioinformatics predictors for classifications", DOI 10.1007/s40484-016-0081-2, 2016.

AUTHORS PROFILE

Ameen Abdullah Qaid Aqlan, received the BCA in Computer Application and M.Sc. in Information System from Osmania University, Hyderabad, Telangana State, India. He is currently Research Scholar in Department of Computer science, Kakatiya University, Warangal, India. His research area is Sentiment Analysis Using Big Data Techniques. He is Members of IAENG and other Professional bodies. He published 3 research papers in international journals.

Dr. B. Manjula, received the BCA in Computer Application and M.Sc. in Information System from Osmania University, Hyderabad, India. She completed the Ph.D in Computer Science from Kakatiya University, Warangal, India. She is currently working as Assistant Professor in Department of Computer Science, Kakatiya University, Warangal, India. She is carrying out her research on Data Analytics, Big Data Analysis, Sentiment Analysis and Optimization Techniques. She has published over 18 research papers international journal and presented more than 15 research papers international conferences. She is a member of ACM, IAENG, CSTA and other Professional Organizations.