International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

A Comparative Study of Twitter Sentiment Analysis using Machine Learning Algorithms in Big Data

C. Bagath Basha, K. Somasundaram

Abstract*: *In recent years, the data growth has suddenly increased through social media like Twitter, Facebook, YouTube, etc., because everyone has used. These unstructured data is used to handle various applications in data analytics. These applications are used for public opinion and sentiment analysis (POSA) in Twitter by various Machine Learning (ML) Algorithms. In this paper, mainly discuss about Twitter sentiment analysis and Machine Learning Algorithms. We take the sample tweets from Twitter, and finding the positive, negative, and neutral words, and then will make it polarity score by using Twitter Sentiment analysis. Using this data are applying ML algorithms. This algorithm is used to show the comparison result between Random Forest (RF) algorithm and Classification algorithm to know which one is best performance. Random Forest algorithm is good when compare with Classification algorithm. Classification algorithm is best for easy understanding. Finally in this social media have low level of security in the Twitter data.

Index Terms*: *Big data, Machine Learning, Twitter

I. INTRODUCTION

In recent times, big data have been increased day by day to meet the needs of social media such as Twitter, Facebook, Google, YouTube and news. During 21st century there is a rapid flow of data. In big data, very large amount of data are moved quickly to various fields over the last 10 years [1, 2]. Additionally, the latest technologies in digital computerized world are unlocked for developing big data [2].

The three main characteristics employed to define the big data are Volume, Velocity, and Variety [3]. Additionally, other characteristics of big data are 5Vs [4], 6Vs [5], and 8Vs [6], where

- In 5Vs stands for Volume, Velocity, Variety, Verification, and Value.
- In 6Vs stands for Volume, Velocity, Variety, Veracity, Visualization, and Value.
- In 8Vs stand for Volume, Velocity, Variety, Veracity, Value, Variability, Viscosity, and Virality.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
C. Bagath Basha*, Research Scholar, Department of CSE, Vinayaka Mission's Research Foundation, Salem, Tamil Nadu, India..
Dr. K. Somasundaram, Professor, Department of CSE, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation, Salem, Tamil Nadu, India..

Volume: It represents the size of data.
Velocity: It indicates the number of times in which the data are produced from various sources.
Variety: It refers to the data generated, from various data formats, such as structure data, semi structure data, and unstructured data.
Veracity: It is to clean the noisy data, to store the cleaned data, and to mine the meaning in data the problem is being analyzed.
Verification: It is to verify quality of data.
Visualization: It is to visualize the data, in any tool and browser without damaging the data i.e. text, image, diagram, etc...
Variability: the data contain a lot of extreme values and they come to statistical problem, and it contains new data.
Viscosity: It helps to understand the elements of velocity.
Virality: It tells how many numbers of users are used and repeated the data from other users.
Values: It is the cost of the big data.

Naturally, big data are of extraordinarily complication in nature at the organization level, and they encounter difficulties in handling and in storing [7]. Hence, we need to improve the method of handling these raw data from a variety of sources [8].Employing in the big data is aimed to analyze the Twitter data with various machine learning algorithms.

II. TWITTER

Twitter is one of the most public social media, it is used to tweet anyone, anywhere, and anytime in the world through internet. 2.1 discusses about the analysis of positive and negative data on Twitter, 2.2 discusses about Machine Learning Algorithms such as Classification, Random Forest, etc.., 2.3 discusses about related studies in Twitter analysis, 2.4 discuss about data collection in Twitter, and 2.5 discusses about related studies in Machine Learning Algorithms.

A. Twitter Sentiment Analysis

Twitter sentiment analyzing for a polarity score in Twitter data. First, we have to find the positive words and symbol them '+'. Second, we have to find the negative words and symbol them '-'. Third, collects the tweets in a particular topic, area, or any other words from Twitter. Fourth, the tweets should be like sentences, in that tweets will make polarity score. Finally, the polarity score should be applying to the various machine learning algorithms.

The positive and negative sample words are below

1. Positive words - **'good, lucky, bless, like, interest, happy'**, etc.
2. Negative words – **'bad, don't, not, sad, won't, bore'**, etc.

Example - tweets,

1. First tweet: "I don't like this movie. It is boring."
2. Second tweet: "I like this movie. It is interesting."

Table 1 gives two simple example tweets of creating the reviews, first tweet *"I don't like this movie. It is boring.* (Class: Negative)*"*, this sentiment the word has two negative words and one positive word, so it is negative tweets. And second tweet *"I like this movie. It is interesting.* (Class: positive)*"*, this sentiment the word has two positive words and no negative word, so it is positive tweets.

Table 1: Analyzing the positive and negative words in tweets for polarity score

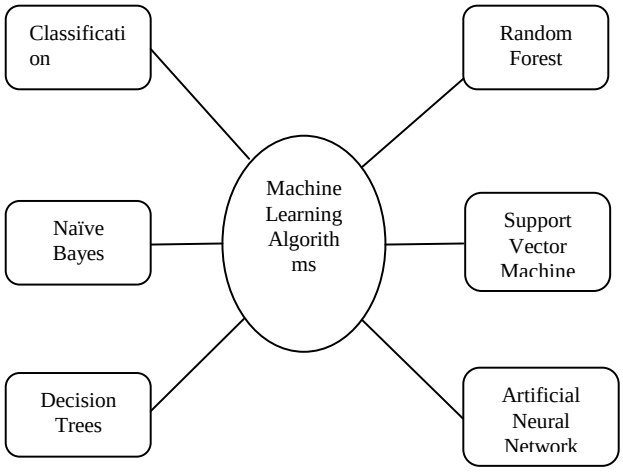
Vocabulary	**Analyzing First tweet positive or negative**	**Analyzing Second tweet positive or negative**
I	1	1
don't	-1	0
Like	+1	+1
This	1	1
movie	1	1
It	1	1
Is	1	1
interesting	0	+1
boring	-1	0
…….		
Class label	Negative (-)	Positive (+)

B. Machine Learning Algorithms

Machine learning algorithms are classified into various types such as Classification, Random Forest, Support Vector Machines, Naïve Bayes, Decision Tree, and Artificial Neural Network.

Fig. 1, **Classification:** It should classify text based and lexicon based on polarity words. This is mainly predicts the sentences Yes, No, and Options that is Positive, Negative and Neutral. Fig. 2, the data should be stored in the x variable, and rpart (Recursive Partitioning and Regression Trees) is a package in Rstudio. To formulate the data, method, actions, are control in rpart to be stored in y. The prune function determines the nested sequence of subtrees with the complexity parameter (cp). The predict is a package in Rstudio, which predicts the data and its class. To make a table of data classification and prediction data is to be stored in t variable. The prob is a probability function in Rstudio, and used to find the probability of t.

Fig. 1 Machine Learning Algorithms



$$
\begin{aligned}
&x = data\\
&y = rpart(x)\\
&y = prune(y, cp)\\
&p = predict(y, type)\\
&t = table(y, p)\\
&prob.table(t) * 100\\
&Note: x\ is\ dataset,\\
&y\ is\ classification\ of\ data,\\
&p\ is\ prediction\ of\ classification.
\end{aligned}
$$

Fig. 2 Classification Algorithm in RStudio

RandomForest: This method has two types of trees such as classification and regression tree. Classification tree is predicting the positive and negative words, and show the result that has a majority of votes. Regression tree also predicts like classification, but it shows the results based on average. Random Forest is mainly predicts the MSE (Mean Square Error) and RMSE (Root Mean Square Error). Fig. 3, the data should be stored in x variable, and set.seed() is to generate the number. To split the two sets of data like 80% and 20% of data from 100% data. Here t_1 is training data and t_2 is test data. To formulate the attribute, data, and mtry is total column number in the Random Forest (rf) is to be stored in y. The predict is a package in Rstudio, which predict the data which is model y and new data. To make a table of training data with prediction data to be stored in mt variable. The prob is a probability function in Rstudio, and find the probability of mt. **Support Vector Machines:** This method is a supervised machine learning algorithm, it can be used both for classifications and regression, but it is mostly used in classification problems, and it is classified into 3 parts that is greatest of middle value in X axis, lowest of middle value in X axis, and average of middle value in Y axis.

Naïve Bayes: This method is a classification algorithm for both two class and multi-class classifications problems. It mainly predicts the probabilities such as class probability and condition probability. Class probability is each class that is positive class or negative class, and Condition probability is every input value in each class.

```
x = dataset
set.seed(1)
id = s(t, p(0.8,0.2))
t1 = x[id ==1, ]
t2 = x[id ==2, ]
attach(t1)
y = rf(attributes, data= t1, mtry= centervalue)
p = predict(y, newdata= t2)
zt2 = t2$attribute - p
mse= mean(zt2^2)
```

Fig. 3 RandomForest Algorithm in RStudio

Decision Tress: It is one of the supervised learning algorithms, and it can be used to predict class and value of target variables from prior datasets. Its representation like tree representation.

Artificial Neural Network: It is based on back propagation algorithm. It is mainly used to reduce the error based on weight value. That weight value should be increased (Forward propagation) or decrease (Backward propagation) the value for error reducing purpose. Fig. 4, the data could be stored in x variable. The apply function is margin of an array, 2 is the start from the second row in dataset, and max is function to store in h variable. Similarly, l variable like h variable. s is a scale function in Rstudio, x is a numeric matrix column, c is center ,either logical value or numeric value, s is a vector of length, and to be stored in z variable. The as.data.frame is function of RStudio, it is checking whether the data is alive, and it's stored in z variable. The attach is a function of RStudio, and this function process the attached data. The neuralnet is a function of RStudio, and it uses backpropagation method to find the hidden values, and stored in y variable. Finally, predict the neural network data and stored in p variable.

C. Related Studies of Twitter Sentiment Analysis

On twitter, opinion mining and sentiment analysis are becoming popular, and have been applied in several application areas such as healthcare, sports, hospitality and tourism, the financial sector, politics, and other related areas which are discussed below.

```
x = dataset
h = apply(x, 2, max)
l = apply(x, 2, min)
z = s(x, c = l, s = h − l)
a(z, 2, h)
a(z, 2, l)
z = as.data.frame(z)
attach(z)
y = neural net(attributes, data = z, hidden = c(2))
p = y$net.result[[1]]
r = attribute − p
mse = mean(r^2)
Note:
h is max, l is min, s is scale, z is scale of data,
a is apply, y is neural data,
p is prediction of neural network,
r is result, mse is mean square root.
```

Fig. 4 Neural Network Algorithm in RStudio

1. Healthcare

Ioannis Korkontzelos et al. [9] proposed adverse drug reaction method (ADR), to show the negative sentiment analysis features. Using ADR post related to 81 drugs that is collected from the DailyStrength forum and Twitter. This method helps to boost the performance in both healthcare related post and tweets in the forum.

Ramon Gouveia Rodrigues et al. [10] this author proposed the sentiment analysis tool for SentiHealth-Cancer (SHC- pt), and this tool detects the positive or negative mood of cancer patients. This proposed tool is compared with other tools such as Seman-tria 3.0.67v, AlchemyAPI 1.1.4v, Textalytics 1.2v, and SentiStrength 0.1.

Erin Hea-Jin Kim et al. [11] focuses on the sentiment dynamics and covers the topic on the hot health issues of Ebola virus in two different media sources such as Twitter and news publications. Ebola virus data were collected from 7106297 tweets and sixteen thousands one hundred and eighty nine news articles from one thousand and six various publication sources respectively.

2. Sports

Robert P. Schumaker et al. [12] this author discussed about central sport system to predict the English Premier League (EPL) matches, and analyzed the sentiment content from last three months of the EPL session. During the 96 hours of tweets, 18027966 tweets were shared using club_hashtag. Before each and every match data were collected from the Twitter API for total of 122 matches. This suggests the crowd-sourced odds and predicts the non positive match's outcome and show tighter goal margin.

Yang Yu and Xiao Wang [13] studied real time sentiment analysis tweets from FIFA World Cup 2014. They collected several tweets from 3 U.S. soccer games and 2 non - U.S soccer game for comparative analysis. The result obtained that was U.S. soccer fans expressed the emotions like fear and anger in tweets, whenever the non-U.S soccer team scored. But the U.S. fans showed more positive tweet emotions such as joy and expectation than the negative emotions during matches. The results were consistent with the predictions of the theory of disposition and showed good and clear predictive validity from fans side.

3. Hospitality and Tourism

Kahlil Philander and YunYing Zhong [14] studied about sentiment analysis for Las Vegas resorts, and the resorts study used the data from Twitter to make the sentiment analysis application at low cost for estimating perceptions of the customer services. A sentiment lexicon methodology was used to create a sentiment index in twitter data. The result shows the performance of comparative analysis for sentiment metrics through Twitter data to know about people's opinion for resorts, and TripAdvisor. This is mainly to concentrate on the experience gained by the hotel customers about facilities and services availed by them.

Rutilio Rodolfo López Barbosa et al. [15] the TripAdvisor gathered the hotel reviews in seven cities, and used to analyze the sentiment and predict the overall hotel ranking through three different algorithms such as a Naïve Bayes algorithm, Boosting and Recursive neural tensor network. The hotel reviews show positive and negative data from the end-users. The usefulness of the three classifiers in the hotel score was compared to the actual scores. It was found that the Boosting and Recursive neural tensor network has better performance when compared with a Naïve Bayesian algorithm.

4. Financial Sector

Thien Hai Nguyen et al. [16] develop a model to predict the stock price model for the sentiment of specific topics of the company. They collected the two types of datasets such as mood information dataset and historical price dataset. These sets of data are employed to evaluate the effectiveness of the model. The SVM is a classifier, and six different methods were applied such as human sentiment, price, sentiment classification, joint sentiment/topic (JST) method, Latent Dirichlet Allocation (LDA) method and aspect based method. The proposed drove stock price model showed the performance with accuracy of about 2% better than that of the above model and prediction of accuracy is about 10% better than the historical price driven model.

Jiajia Li and Phayung Meesad [17] proposed the model of sentiment analysis based on prediction model in the inverse bias algorithm. And also proposed another algorithm which is a semi supervised naïve bayes classification algorithm. These algorithms are based on the SVM linear algorithm with hybrid features. This experiment of 4622 tweets data collected from Topsy.com. Using these algorithms the accuracy increased from 86.95% to 90.33%.

Wenhao Chen et al. [18] according to him people voiced their opinion or emotions related to the stock market of china. The number of users were collected the data from social media websites. This authors used a dynamic programming in Chinese segmentation tool Jieba, and categorizing the word from text into different emotion categories in Chinese Emotion Word Ontology. This emotional state of discussing and happiness strongly predict the stock market price in china.

5. Politics

Andrea Ceron et al. [19] this author proposed method called Hopkins and King, and they have two tracking the process. The first tracking process involves checking the online popularity of different Italian politics till 2011, and second tracking process involves French people choosing the candidate for President in 2012. The choosing of presidential candidates and the results were predicted through social media.

Cesar Alfaro et al. [20] the main purpose of this paper is the natural recognition of several different opinion and trend in the weblogs for the multi-stage method, and other purpose is supervised and unsupervised techniques in machine learning algorithm for opinion mining and sentiment analysis. These techniques are used to analyze the feedback of each and every company products and services through social media. The test result show's both Support Vector Machine (SVM) classification and KNN for increasing in the accuracy.

6. Other application areas

Shahid Shayaa et al. [2] this author mainly focuses on the technical aspect and non-technical aspect. The technical aspect is OMSA technical and non-technical aspect is various applications. Both these techniques were used for analysis of the literature survey for future direction in research.

Jinsong Wu et al. [21], in this author has studied about the processing the enormous data and the processing of big data life cycles like data generation, data acquisition, data communications, data storage, data analysis and processing. They proposed two new metrics which are effective energy efficient and effective resources efficient (2016).

Doug Laney [22] META Group in 2001 proposed for the individual of big data, with 3 Vs, i.e., Volume, Velocity, and Variety. Bharath Sriram et al. [23] in his work classified the tweets to a set of generic classes, including event, opinions, private messages, news, and etc.. However, most of the works are mainly focused on the content of the tweets and how opinions of users are extracted towards specific objects or topics. Dave Beulke [4], this author has proposed 5 Vs of big data, i.e. Volume, Variety, Velocity, Veracity, and Value, has been discussed previously in 2011.

Enterprise Strategy Group [5], they proposes for the 6 Vs of big data, i.e. Volume, Variety, Velocity, Veracity, Visualization, and Values, has been introduced in 2012.

William Vorhies [6], the authors proposed in 2014 about the 8 Vs of big data i.e., Volume, Variety, Velocity, Veracity, Value, Variability, Viscosity, and Virality.

Masahiko Itoh et al. [24] author has studied and proposes two forms of big data in visual integration of traffic analysis and social media analysis for smart card data on the Tokyo Metro and social media data on Twitter. A behavior and abnormal situations are extracted from smart card data and reflects real situations such as disasters, accidents, and public gatherings.

There are three visualization components in the analysis of system such as HeatMap view, AnimatedRibbon view, and TweetBubble view.

D. Data collection in Twitter

Table 3: The Dataset in Sentiment Classification

Dataset	#Positive	#Negative	#Neutral	Total Dataset
RAJINI	519	4	477	1000
AKSHAY	360	2	638	1000
DHONI	538	97	365	1000
KOHLI	493	25	482	1000

Table 3, it shows the various famous peoples positive, negative, and neutral data set in sentiment classification.

E. Related studies in Machine Learning

Machine Learning is an ability to learn automatically and improve the performance to make better decision in future based on prediction. In this chapter we discuss about the advantage and disadvantages of Machine learning algorithms.

1. Artificial Neural Network (ANN)

ANN is a mathematical model approach that is developed in the effective process of the human mind [25]. This method performs data modeling, which determines whether it is a supervised learning technique or unsupervised learning technique. The supervised learning provides both the input and output data, ANN use this input to generate the output. Unsupervised learning provides only input data and ANN is used to find the output data.

2. Random Forest (RF)

RF is a classification and regression method [26], which can create a number of classifiers and produce an aggregate result [27]. Two methods are used for the classification of trees such as boosting [36] and bagging [28].

3. Support Vector Method (SVM)

Support vector method is a binary classification as well as multiclass classification [2]. In multiclass SVM is the number of binary classifiers are fabricating and conjoining, they directly maintain all the data in one optimization construction [29].

4. Genetic Algorithm (GA)

Genetic algorithms are optimization techniques and having many different search spaces, the concept of genetics is used theatrically to create an algorithm which is robust, efficient, and flexible in nature [2].

5. Naïve Bayes (NB)

The Naïve Bayes classifier is a supervised machine learning technique, and it is based on the theorem of probability. This algorithm is applying to numeric data [30], which is understand easily, quick, and simple for classification [2].

6. Decision Tree (DT)

The main purpose of DT classifier is predict and classify the tasks. It is very easy to create the decision tree rules, which are given in the hierarchical representation. The tree is composed of decision along with the event nodes, path, and edge [31].

7. Keyword based classification (KBC)

The KBC method mainly classifies the text based on the polarity words such as joyful, happy, very happy, sad, and very sad [32]. The main drawbacks of the KBC is unable to identify the proper negative words [32].

8. Lexicon based classification

This method creates a list of words which are manually labeled as positive and negative polarity, and creates a polarity score for each word. The main advantage of this method does not need training data. This method is useful in conventional text such as reviews, forums, and blogs [2, 33].

III. COMPARISON OF VARIOUS SENTIMENT ANALYSES IN TWITTER

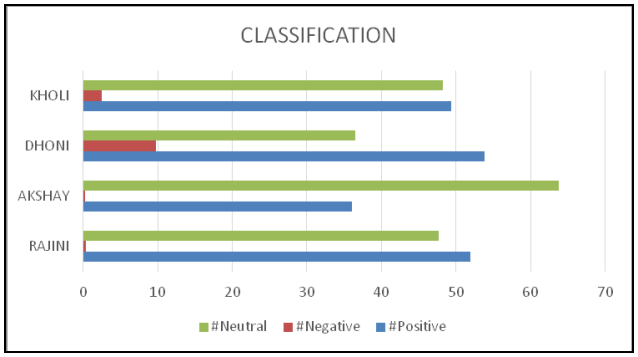


Fig. 5 Comparison of Public Opinion on Twitter

Fig. 5, shows the general public opinions on Twitter for famous people kohli has positive and neutral opinions are more or less same in the given graph, but comparing with others it has a third level in the graph. Dhoni has more positive opinion and more negative opinion than others

in the Twitter sentiment analysis, and also neutral opinion very less when compared to others in the given graph. Akshay has a more neutral public opinion and more less

Table 2: Areas of Big data and sentiment analysis

S. No.	Authors	Purpose	Data sets	Area
Health Care				
1	Ioannis et al. [9]	ADR method is to search and process the sentiment analysis features.	Twitter	Drugs
2	Ramon et al. [10]	To detect the positive or negative mood of cancer patient through online.	Face-book	Medic-al Science
3	Erin et al. [11]	Two different media sources are investigated the Ebola virus.	Twitter and News Publications	Health Care

Sports				
4	Robert et al. [12]	To analyze the emotions tweets and predict in English Premier League matches.	Twitter	Sports
5	Yang Yu et al. [13]	To analyze the U.S. Soccer Fans emotions through tweets during the match.	Twitter	Sports
Hospitality				
6	Kahlil et al. [14]	To analyze the real time hospitality and customer reviews through the Twitter sentiment analysis.	Twitter	Hospitality and Tourism
7	Rutilio et al. [15]	Trip-advisor analyzes the hotel reviews using sentiment to analyze algorithm.	Tripadvisor.com	Hospitality
Financial Sector				
8	Thien et al. [16]	Develop a model to predict the stock price for the sentiment on the specific topics of the company	Yahoo finance message board	Stock Market
9	Jiajia et al. [17]	To predict the stock market trends in social media.	Topsy.com	Stock Market
10	Wenhao et al. [18]	To analyze the public opinion and emotions of the stock market in China.	China website	Stock Market
Politics				
11	Andrea et al. [19]	To analyzed the political preference data in Twitter.	Twitter	General Public
12	Cesar et al. [20]	To analyze the every company through supervised and unsupervised machine learning algorithm.	Weblogs	Voters
Other Related				
13	Shahid et al. [2]	To analyze the various applications and algorithms for literature purpose.	Social Media	Big Data
14	Jinsong Wu et al. [21]	To analyze the large data and process of big data life cycles.	Social Media	Big Data
15	Doug Laney [22]	Proposed the 3Vs	Social Media	Big Data
16	Dave Beulke and Associates [4]	Proposed the 5Vs	Social Media	Big Data
17	Enterprise Strategy Group [5]	Proposed the 6Vs	Social Media	Big Data
18	William Vorhies [6]	Proposed the 8Vs	Social Media	Big Data
19	Masahiko et al. [24]	To analyze the traffic through social media on Tokyo Metro	Social Media	Big Data
20	Bharath et al. [23]	To analyze the tweets for public opinion on Twitter	Twitter	Twitter

Table 4. Advantage and disadvantage of machine learning algorithms

S. No.	Machine Learning Algorithms	Working Principle	Advantage	Dis Advantage
1	Artificial Neural Network (ANN)	The main purpose of ANN is information processing and Knowledge representation. [34]	The algorithm gives better accuracy in several sentiment applications.	Need more processing power and graphical processing units.
2	Random Forest (RF)	Is a classification and regression method based on a group of amount of decision tree [26].	Is a good algorithm used for complex classification tasks, and that model created can be easily interrupted.	This algorithm is slow and ineffective for real time predictions in large amount of data.

3	Support Vector Machine (SVM)	The form of radical basic function and it is used to learn in different applications, and used for suitable kernel function [35]	It is a very good method, and work with unstructured and semi-structured data. The risk of over-fitting is very less.	It is not easy to choose good kernel function.
4	Genetic Algorithm	These are optimization techniques having many different search spaces, and the concept of genetics is artificially used to create an algorithm which is robust, efficient, and flexible in nature [2]	It's very easy to understand the concept of genetic algorithms, and optimization also good for noisy environment.	It's not safe to finding data in a particular period.
5	Naïve Bayes (NB)	It is mainly based on probability theorem, and this algorithm is applying to numeric data [30].	It's easy and simple understood for classification, and require less training data.	NB is to reduce the effectiveness of a system.
6	Decision Tree	It is very easy to understand for creating the decision tree rules, and is given in the hierarchical representation. The tree is composed of decision & event nodes, path, and edge [31].	It is very easy to understand and interrupt, and transparency, specificity, and comprehensive nature.	It is unstable in complexity and calculation part.

negative public opinion while comparing to others, and positive opinion also good at Twitter. Rajini has second most positive opinion in the graph, neutral opinion also good while comparing to others, and negative opinion is very less in Twitter.

Fig. 6, Kohli has positive and neutral opinions are same, and also have negative opinions. Dhoni has a more positive opinion and more negative opinion than others in Twitter, and also neutral opinions are good to compare with others in the given graph. Akshay has a more neutral public opinion and less negative public opinion while comparing to others, and positive opinion is good in Twitter. Rajini has second most positive opinion and neutral opinion also good while comparing to others, and negative opinion is more on Twitter.

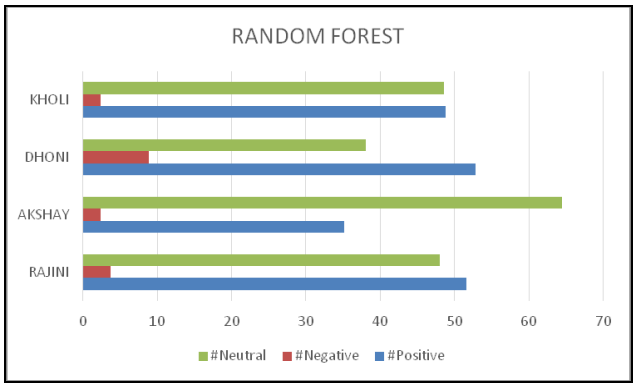


Fig. 6 Comparison of public Opinion on Twitter

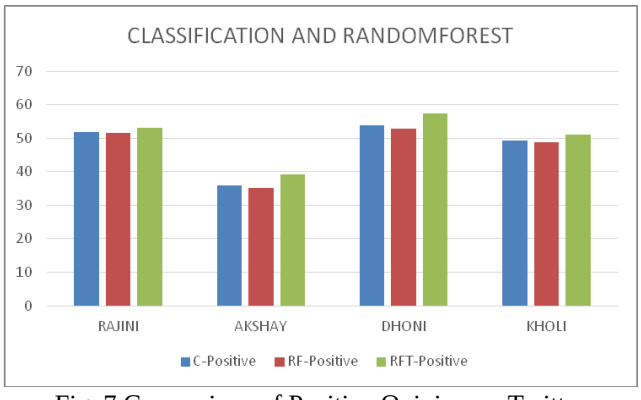


Fig. 7 Comparison of Positive Opinion on Twitter

Fig. 7, C- classification, RF- RandomForest, RFT-RandomForest test algorithm shows the comparison of positive opinions. These algorithms used to show more positive opinions while compared to others, and all are having almost the same result, but RFT positive opinions are more comparable to others.

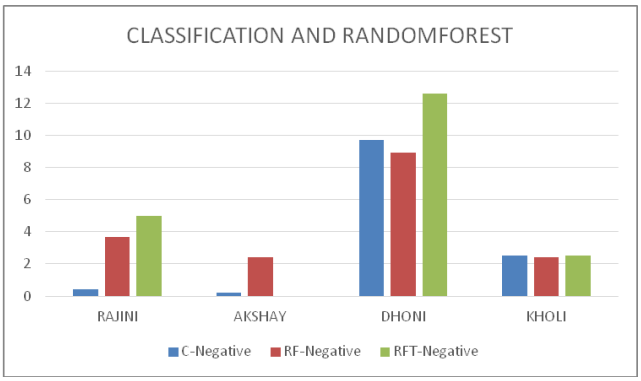


Fig. 8 Comparison of Negative Opinion on Twitter

Fig. 8, C- classification, RF- RandomForest, RFT-RandomForest test algorithm show the comparison of negative opinions. These algorithms show's the result Dhoni have more negative opinions while compared to others. Rajini has less negative opinions through classification algorithm, and randomforest algorithm shows more negative opinions on Twitter. Akshay has very less negative opinions when compared to others in the classification, and randomforest also showing a less negative opinion to compare with others. Kohli has negative opinions and more over same in these algorithms.

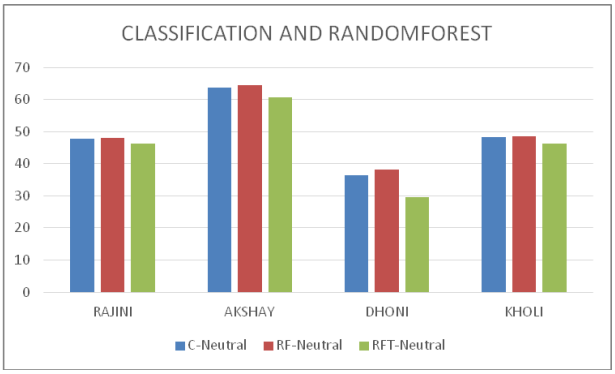


Fig. 9. Comparison of Neutral Opinion in Twitter

Fig. 9, C- classification, RF- RandomForest, RFT-RandomForest test algorithm show the comparison of neutral opinions. These algorithms showed the result between AKSHAY and other three are having good neutral opinions, but Akshay has more neutral opinions while compared to others. Rajini and Akshay have more over the same range, but Dhoni has less opinions while compared to others.

IV. DISCUSS AND CONCLUSION

Today, it is very important to have social media to analyze the sentiment of people opinions by having reviews, Facebook, forum blogs, 80% of peoples monitoring the social media data for analysis purpose. For example, every tweet has a maximum of one hundred and forty characters. These tweets can assign a polarity scores through the software. That scores to evaluate whether the tweet looks like positive or negative or neutral. These types of research already done in various applications or methods of public opinion and sentiment analysis. In addition, the applications of public opinion and sentiment analysis for stars, government and politics views are still growing. Here we take sample data from Twitter and used this data to compare two algorithms such as Classification and RandomForest algorithms. The comparison result shows RandomForst algorithm is good when compared with classification, but classification is very easy to understand and implementation against RandomForest. Finally, it is used to share the new product and government information's to reach quickly for public or users. Finally, all data are unsecured because public opinions only show the result between good or bad. In future we need security of these social media data.

REFERENCES

1. Richard Addo-Tenkorang and Petri T. Helo, ``Big data applications in operations/supply-chain management: A literature review", Computers & Industrial Engineering., vol. 101, Nov. 2016, pp. 528-543.
2. Shahid Shayaa, Noor Ismawati Jaafar, Shamshul Bahri, Ainin Sulaiman,Phoong Seuk Wai, Yeong Wai Chung, Arsalan Zahid Piprani, and Mohammed Ali Al-Garadi, "Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges", IEEE Access, vol. 6, 2018, pp. 37807-37827.
3. Paul C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch, and George Lapis, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data", New York, USA: McGraw-Hill, 2011.
4. Dave Beulke (2011, November), "Big data impacts data management: The 5 Vs of big data". Available: http://davebeulke.com/big-data-impacts-data-management-the-five-vs-of-big-data/.
5. Enterprise Strategy Group (2012, August), "The 6 Vs: The BI/analytics game changes so Microsoft changes excel". Available: http://www.esg-global.com/blogs/the-6-vs-the-bianalytics-gamechanges-so-microsoft-changes-excel/.
6. William Vorhies (2014, October), "How many V's in big data? The characteristics that define big data". Available: http://www.datasciencecentral.com/profiles/blogs/how-many-vs-in-big-data-the-characteristics-that-define- big-data.
7. Anuranjan Misra, Anshul Sharma, Preeti Gulia, and Akanksha Bana, "Big data: Challenges and opportunities" International Journal of Innovative Technology and Exploring Engineering. vol. 4, 2014, pp. 41-42.
8. M. Batty, K.W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart cities of the future", Eur. Phys. J. Special Topics, vol. 214, no. 1, Nov. 2012, pp. 481-518.
9. Ioannis Korkontzelos, Azadeh. Nikfarjam, Matthew Shardlow, Abeed Sarker, Sophia Ananiadou, and Graciela H. Gonzalez, "Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts", Journal of Biomedical Informatics, vol. 62, Aug. 2016, pp. 148-158.
10. Ramon Gouveia Rodrigues, Rafael Marques das Dores, Celso G. Camilo-Junior, and Thierson. Couto Rosa, "SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks", International Journal of Medical Informatics, vol. 85, 2016, pp. 80-95.
11. Erin Hea-Jin Kim, Yoo Kyung Jeong, Yuyoung Kim, Keun Young Kang, and Min Song, "Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news", Journal of Information Science, vol. 42, 2016, pp. 763-781.
12. Robert P. Schumaker, A. Tomasz Jarmoszko, and Chester S. Labedz Jr , "Predicting wins and spread in the Premier League using a sentiment analysis of Twitter", Decision Support Systems, vol. 88, Aug. 2016, pp. 76-84.
13. Yang Yu and Xiao Wang, "World cup 2014 in the Twitter world: A big data analysis of sentiments in US sports fans' tweets", Computers in Human Behavior, vol. 48, Jul. 2015, pp. 392-400.
14. Kahlil Philander and YunYing Zhong, "Twitter sentiment analysis: Capturing sentiment from integrated resort tweets", International Journal of Hospitality Management, vol. 55, May 2016, pp. 16-24.

15. Rutilio Rodolfo López Barbosa, Salvador Sánchez-Alonso, and Miguel Angel Sicilia-Urban, “Evaluating hotels rating prediction based on sentiment analysis services”, Aslib Journal of Information Management, vol. 67, 2015, pp. 392-407
16. Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin, “Sentiment analysis on social media for stock movement prediction”, Journal of Expert Systems with Applications, vol. 42, 2015, pp. 9603-9611.
17. Jiajia Li and Phayung Meesad, “Combining sentiment analysis with socialization bias in social networks for stock market trend prediction”, International Journal of Computational Intelligence and Applications, vol. 15, 2016, pp. 1650003-1 - 1650003-16.
18. Wenhao Chen, Yi Cai, Kinkeung Lai, and Haoran Xie, “A topic-based sentiment analysis model to predict stock market price movement using Weibo mood”, Web Intelligence, vol. 14, 2016, pp. 287-300.
19. Andrea Ceron, Luigi Curini, Stefano M Iacus, and Giuseppe Porro, “Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France”, New Media Society, vol. 16, 2014, pp. 340-358.
20. Cesar Alfaro, Javier Cano-Montero, Javier Gómez, Javier M. Moguerza, and Felipe Ortega, “A multi-stage method for content classification and opinion mining on weblog comments”,' Ann. Oper. Res., vol. 236, 2016, pp. 197-213.
21. Jinsong Wu, Song Guo, Jie Li, and Deze Zeng, “Big Data Meet Green Challenges: Greening Big Dat”, IEEE Systems Journal, Vol. 10, September 2016, pp. 873-887.
22. Doug Laney, “3D data management: Controlling data volume, velocity, and variety”, Application Delivery Strategies, META Group, Feb. 2001.
23. Bharath Sriram, David Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas, “Short text classification in Twitter to improve information filtering”, in Proceeding. 33rd International. ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 2010, pp. 841–842.
24. Masahiko Itoh, Daisaku Yokoyama, Masashi Toyoda, Yoshimitsu Tomita, Satoshi Kawamura, and Masaru Kitsuregawa, “Visual Exploration of Changes in Passenger Flows and Tweets on Mega-City Metro Network”, IEEE Transactions on Big Data, Vol. 2, 2016, pp. 85-99,.
25. Warren S. McCulloch and Walter Pitts, “A logical calculus of the ideas immanent in nervous activity”, Bulletin of Mathematical Biology, vol. 5, pp. 115-133, 1943.
26. Anne-Laure Boulesteix, Silke Janitza, Jochen Kruppa, and Inke R. König, “Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics”,' Wires Data Mining & Knowledge Discovery, vol. 2, 2012, pp. 493-507.
27. Andy Liaw and Matthew Wiener, “Classification and regression by randomforest”, R News, vol. 2, 2002, pp. 18-22.
28. Leo Breiman, “Bagging predictors”, Machine Learning, vol. 24, 1996, pp. 123-140.
29. Chih-Wei Hsu and Chih-Jen Lin, “A comparison of methods for multiclass support vector machines”, IEEE TRANSACTIONS ON NEURAL NETWORKS, vol. 13, Mar. 2002, pp. 415-425.
30. Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz, “A Bayesian approach to filtering junk e-mail”, in Proceeding, AAAI Workshop on Learning for Text Categorization, AAAI Technical Report, vol. 62, 1998, pp. 55-62.
31. J. R. Quinlan, “Induction of decision trees”, Machine Learning, vol. 1, 1986, pp. 81-106.
32. Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi, “New avenues in opinion mining and sentiment analysis”, IEEE Intelligent Systems., vol. 28, no. 2, Mar. 2013, , pp. 15-21.
33. Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede, “Lexicon based methods for sentiment analysis” Computational Linguistics, vol. 37, pp. 267-307, 2011.
34. Enso Grossi and Massimo Buscema, “Introduction to arti_cial neural networks”, European Journal of Gastroenterology & Hepatology, vol. 19, Dec. 2007, pp. 1046-1054.
35. Thorsten Joachims, “Text categorization with support vector machines: Learning with many relevant features”, Machine Learning: ECML, vol. 98, 1998, pp. 137-142.
36. Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee, “Boosting the margin: A new explanation for the effectiveness of voting methods”, The Annals of Statistics, vol. 26, 1998, pp. 1651-1686.

AUTHORS PROFILE

C. Bagath Basha is having teaching experience about 6 years and 6 months. He served in various positions in Teaching. He is currently doing as Research Scholar, Department of Computer Science and Engineering, Vinayaka Mission’s Research Foundation, Salem, Tamil Nadu, India. His area of interest includes Big Data and Data Analytics, Security.

Dr.K.Somasundaram is having industry and teaching experience about 24 years. He served in various positions in industry and Teaching. He is currently serving as Professor and Program Director (Engineering Research) in Computer Science and Engineering department at Aarupadai Veedu Institute of Technology, Vinayaka Mission’s Research Foundation, Chennai. He published about 80 papers in International journals and presented 32 papers in refereed national & International Conferences. There are 8 scholars are completed their research under his guidance and 8 PhD scholars are doing their research. He guided more than 23 M.E., Thesis. He is a member of IE(India), IETE,CSI, ISTE and CEng(IEI). His area of interest includes Data Mining and Data Analytics, Wireless Sensor Networks, Grid/Cloud computing.