Sentiment Analysis Using Supervised Machine Learning Algorithms

Niharica choubey[1], Vilender Kumar[2], Sanjay Kumar Gupta[3]
M.Phil student, School of studies in Computer Science & Applications
Jiwaji University Gwalior (M.P.), India[1]
Associate Professor, Gitarattan International Business School, New Delhi 110085[2]
Professor, School of studies in Computer Science & Applications
Jiwaji University Gwalior (M.P.), India[3]

Abstract- *Today, social media became a big platform of data collection. People of all over the world give their own opinion on social sites. Therefore, datasets are collected from social media platform twitter regarding three Indian politicians as tweets and, three supervised machine learning algorithms as Support Vector Machine (SVM) with four kernel, Random Forest, and Naive Bayes are used for analysis of six sentiments for prediction of popularity in election 2019. Furthermore, comparisons between three algorithms are also performed.*

Keywords- Machine learning algorithm, SVM, Random Forest, Naive bayes, Text Mining, Indian politicians

I. INTRODUCTION

Today, the popularity of social media is increasing very fast in many ways. A person of every age like youngsters, elders, and children are utilizing social media platforms and gives their opinion about various topics such as feedback of products, movie review and many others. Online sites or social media applications gathered more information or data which are given by social media users. So, new researcher's interest is increasing in the sentiment analysis or opinion mining [1]. However, Artificial intelligence gives a big contribution of developing a sentiment analysis. Artificial intelligence is the process of making an intelligent machine which acts and thinks like a human being. When two persons communicate to each other; so in this case, one person feelings are easily understandable by other person, but when person communicate with machine then how will machine understand person opinion, words etc. When person gives the opinion, share own feelings on social media and people gives feedback about products, then machine can easily calculate sentiments of people which are positive, negative, neutral, sad, happy using many machine learning algorithms. Many algorithms have been designed for sentiment analysis such as supervised algorithm, unsupervised algorithm, reinforcement algorithm, semi-supervised algorithm. These algorithms are part of machine learning algorithm.

User of social media network is increasing on Twitter, and generate about 500 million tweets per day which allows people to share their thinking and feelings rapidly and spontaneously. Large amount of usage of Tweeter reflects the need of analyzing sentiments expressed in tweets to identify and extract public emotions from text messages for accurate and automatic sentiment classification to predict popularity in political election 2019.

Thus, objective of this work is to analyze opinion of public as sentiments regarding two Indian politicians using three supervised machine learning algorithms and tweets as social media posts collected from twitter from 2014 to 2018 to analyze the popularity for election 2019. Finally, predict popular Indian leader in public of 2019 elections and also predict the best algorithm for sentiment analysis. However, dataset size used in this work needs large data and more collection to set the precise and accurate trends. It needs more extensive perfection based on large social media posts of tweeter as datasets using supervised machine learning algorithms with clear-cut purpose to ascertain to analysis of six sentiments.

The outline of this paper is as follows. We start with significance of sentiment analysis in section 1. Section 2, contains the related work done by researchers and scientists for sentiment analysis across the globe. In section 3, provides methodology using supervised machine learning algorithms for sentiment analysis. Results of supervised learning algorithms are illustrated in section 4. Summarized outcomes are further discussed in section 5. Lastly, conclude in section 6.

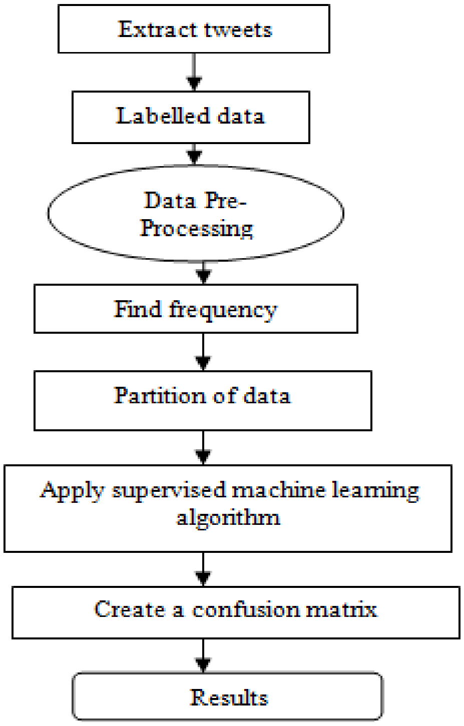
II. RELATED WORK

Many research works have done on machine learning algorithm. D. Krishna madhuri [2] explained case study on Indian railway. Datasets are collected from twitter that analyse people opinion about Indian Railway that means, what people think about Indian Railway. Three types of sentiments are described positive, negative, neutral using machine learning framework. It explains four techniques of machine learning framework like c.4.5, naive bayes, support vector machine, random forest. In this paper, accuracy, recall, precision, F-measure results are compared and also calculated false positive rate. Erik boiy, marie-francine moens [3] described collection of tweets on three languages English, Dutch, French from blog, review, forum text. It expresses people feelings on different products in different languages. Several supervised classification method are used like Multinomial Naive bayes, SVM, Maximum Entropy. It gains three sentiments positive, negative, and neutral. This research paper, increase accuracy of classifiers using cascade rule, active learning and many more results are found. In this study [4], researcher analyzes sentiments of public about Pakistan political parties. Collections of tweets are in Urdu, English, roman Urdu language from twitter. To get polarity (positive, negative, neutral) and subjectivity calculator used sentiment analyzers like textblob, sentiwordnet, and word sense disambiguation sentiment analyzer with two supervised machine learning algorithms such as naive bayes and support vector machine. In this research, comparison of three lexicons based sentiment analyzers was used. In [16], Tidytext mining technique is used for sentiment classification of twitter data. Three Lexicon based methodologies NRC, BING, and AFINN are utilized to analyse emotions for sentiment analysis to predict popular Indian leader in election 2019.

III. PROPOSED METHODOLOGY

Our study has done on sentiment analysis of twitter data using supervised machine learning algorithm. Before using these algorithms on dataset we face many problems during experiment like when we collects data from social sites the data is in unstructured or unreliable format and many words are exist in dataset which is useless. So, first we have to remove these words and make the dataset in reliable format. I have used R language for all experiment work. This section explores into four parts A. Data gathering B. Data-pre-processing C. Machine learning algorithm.

Workflow diagram



Part1: Data gathering

Data is collected from twitter social media platform where it have huge amount of data about politics and politician. So, three Indian leaders Sh. Narendra Modi, Sh. Akhilesh Yadav, and Sh. Rahul Gandhi are considered, and for this 2,500 tweets are collected of each politician. twitteR package and twitter API are used for collecting tweets. For finding popular politician two dataset of 2500 each are used while compare algorithm 7,500 dataset is used which is a combination of three dataset. After extracting the data, dataset is labelled in six classes or emotions. These classes are positive, negative, anger, disgust, happy and neutral.

Part 2: Data Pre-processing

After collection of tweets, data pre-processing is very important for best result. In Data pre-processing, some unreliable information, character, tags like # tags, punctuation marks, stop words etc are removed [5,16]. For this, text mining package is used which is given by R language.

Text mining: - data pre-processing work easily done with text mining (tm package) because text mining package gives already made function which can easily remove unreliable data from dataset. Dataset summarize by text mining process. We can say in easy words, text mining techniques convert data in easy format that the users understand whether the information is useful or not [6]. Before using text mining, first text data is converted into a factors using function as.fatcor ( ). When we analyze the text in R language, it is very important to convert your text data into Unicode standard with 8 –bit block Unicode transfer format UTF-8 using iconv ( ) function [7]. Now, create a corpus by text mining package. Corpus convert data frame into a document that means each sentence convert into documents therefore we can say that the corpus is a collection of documents. After that, cleaning of text data starts using various functions like RemovePunctuation- removes punctuation marks from dataset like #, @, comma etc.

RemoveNumbers- removes unwanted numbers which is written between texts.

Tolower- converts text data into a lower case from upper case.

Remove Stop words- remove stop words which is written in to text like the, about, of, in, another and many more.

Remove URL links- remove hyperlinks.

Remove extra space- when we remove the punctuation, stop words, numbers from text so the space is empty between sentences. This space is removing using stripWhitespace ( ) function.

Before cleaning text

[1] "We are banker we want justice... We demand ops (old pension scheme), 5 days banking, better and speedy wage settlem… https://t.co/xAHKqyLUvL"

[2] "RT @drraghu12345: @iPuneetSharma @hindustanse @RahulGandhi @narendramodi @AmitShahOffice @PiyushGoyal @nikhildadhich @SonunigamsingH @ihite…"

[3] "@hellomanish_kr @BJP4India @BJP4UP @BJPLive @dr_maheshsharma @narendramodi @myogiadityanath @FightForRERAInd… https://t.co/YIB4GgEAG5"

After cleaning text
<<SimpleCorpus>>
Metadata: corpus specific: 1, document level (indexed): 0
Content: documents: 5

[1] banker want justice demand ops old pension scheme days banking better speedy wage settlem…

[2] drraghu ipuneetsharma hindustanse rahulgandhi narendramodi amitshahoffice piyushgoyal nikhildadhich sonunigamsingh ihite…

[3] hellomanishkr bjpindia bjpup bjplive drmaheshsharma narendramodi myogiadityanath fightforreraind…

After clean the dataset, text is converted into a document term matrix. Terms and sparsity are generated from documents using the function DocumentTermMatrix ( ). Here, also found non-spars entry and spars entries. Spars entries can also be removed from term using remove sparse term function.

Part3: Machine learning algorithm

Supervised machine learning algorithms are used for classification. In this work, six types of emotions as happy, positive, negative, anger, disgust, and neutral are analyzed. Machine learning algorithms work on dataset, analyse the dataset and give the response. The training and testing processes are used by Machine learning algorithms and produce the results. Supervised algorithms use training and testing process and include many classifiers like Probabilistic classifier, Naive bayes classifier, Tree classifier etc [8]. In this, three supervised algorithms; Support Vector Machine (SVM), Naive bayes, and Random Forest are used.

A. Support Vector Machine (SVM): - SVM algorithm produces best result when we classify sentiment analysis [9]. When data is analyzed for classification then SVM find hyperplane between two classes. It is used for data classification and regression but mostly used for classification [10] [11]. When we work on linear dataset then there is no problem for finding a hyperplane but when we work on non-linear dataset, then we face problems for classification. Therefore kernel trick of SVM is used because some problems occur when hyperplane is found on non-linear dataset. Combinations of mathematical functions are known as kernel. Kernel takes information as an input and after that transform into a required form [11]. Many kernel functions are as follows: linear, non-linear- polynomial, sigmoid, radial basis function, Fisher kernel, string kernel, graph kernel [12]. The kernel is used during training process and testing process for predicting the results [7]. But, in this paper, only four kernel linear, non-linear- polynomial, sigmoid, radial basis functions are used. Mathematical formula of these kernels are :

1. Linear- u`* v
2. Polynomial- (gamma*u'*v + coef0)^degree
3. Radial basis- exp(-gamma*|u-v|^2)
4. Sigmoid- tanh (gamma*u'*v + coef0) [7].

B. Naive Bayes: - Naive Bayes classifier is type of supervised machine learning algorithm. Naive Bayes algorithm depends on probability of words. Its mathematical calculation based on Bayes theorem formula which depends on conditional probability [9] [10] [13].

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

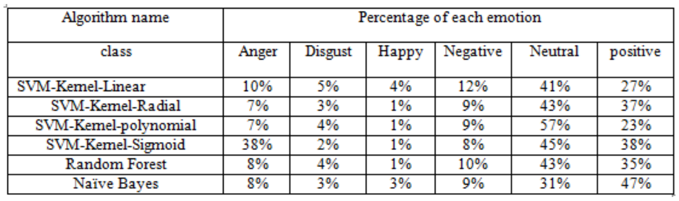
C. Random forest algorithm: - Random forest classifier used for regression and classification like support vector machine.

Random forest algorithm works like a decision tree algorithm. This algorithm is based on tree-based methods. It creates a forest with a number of trees, where each tree grows in ensemble according to random parameter [14]. This algorithm creates a many decision trees from dataset and gets the prediction result from each of them and selects the best solution from majority voting [15].

IV. EXPERIMENT AND RESULTS

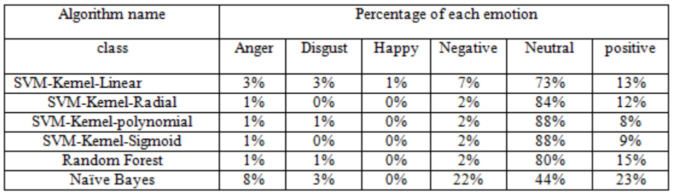
In this section, three algorithms are applied on 2,500 dataset of Sh. Rahul Gandhi and Sh. Narendra Modi and found the popular Indian leader in 2019 elections. For this work, partitioned of dataset in two sections. First section take data from dataset for training which is 80% and second section take data from dataset for testing which is 20%. Sentiment classification performs on six classes Neutral, positive, negative, anger, happy, disgust.

Table1. Percentage table of each class on Sh. Narendra Modi dataset



Algorithm name	Percentage of each emotion					
class	Anger	Disgust	Happy	Negative	Neutral	positive
SVM-Kemel-Linear	10%	5%	4%	12%	41%	27%
SVM-Kemel-Radial	7%	3%	1%	9%	43%	37%
SVM-Kemel-polynomial	7%	4%	1%	9%	57%	23%
SVM-Kemel-Sigmoid	38%	2%	1%	8%	45%	38%
Random Forest	8%	4%	1%	10%	43%	35%
Naïve Bayes	8%	3%	3%	9%	31%	47%

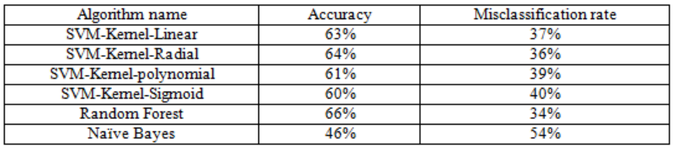
Table2. Percentage table of each class on Sh. Rahul Gandhi dataset



Algorithm name	Percentage of each emotion					
class	Anger	Disgust	Happy	Negative	Neutral	positive
SVM-Kemel-Linear	3%	3%	1%	7%	73%	13%
SVM-Kemel-Radial	1%	0%	0%	2%	84%	12%
SVM-Kemel-polynomial	1%	1%	0%	2%	88%	8%
SVM-Kemel-Sigmoid	1%	0%	0%	2%	88%	9%
Random Forest	1%	1%	0%	2%	80%	15%
Naïve Bayes	8%	3%	0%	22%	44%	23%

In above table-1 and table-2, the positivity and happiness is high of Sh. Narendra Modi in comparison to Sh. Rahul Gandhi. So, it can say that Sh. Narendra Modi is famous leader in 2019 election. Now, average accuracy of three algorithms is calculated. For this, combined three dataset are used which was collected from twitter as tweets from 2014 to 2018, thus total dataset size is 7,500.

Table-3. Accuracy



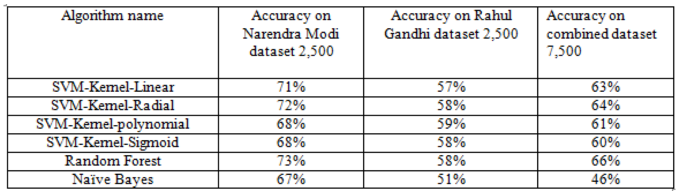
Algorithm name	Accuracy	Misclassification rate
SVM-Kemel-Linear	63%	37%
SVM-Kemel-Radial	64%	36%
SVM-Kemel-polynomial	61%	39%
SVM-Kemel-Sigmoid	60%	40%
Random Forest	66%	34%
Naïve Bayes	46%	54%

Table-3 shows average accuracy and misclassification rate of algorithms. When algorithm performs prediction on dataset which is based on training process then find out how many data are misclassify by classifier. More misclassification rate means low accuracy and low misclassification rate indicates high accuracy. So, highest accuracy is produced by random forest algorithm with low misclassification rate as in table-3. If SVM four kernels is compared then the radial basis kernel function produced highest accuracy 64% than other kernel. Naive bayes performed very low accuracy with high misclassification rate. Thus, it can say that results produced in table-1 and table-2 by Random Forest algorithm is more accurate rather than others.

V. DISCUSSION

In this work, sentiment analysis is performed on twitter data using supervised machine learning algorithms. Three algorithms SVM with four kernel, Naive bayes, and Random Forest are used for this work. Four kernels of SVM and two other algorithms are compared. In section 4, accuracy and misclassification rate of these algorithms are calculated, and thus Random Forest algorithm is having highest accuracy among all. But when we run Random Forest algorithm some problems occur like it takes more time than other algorithms. Random Forest training time based on their own trees which means as we will use more trees then it will take more time. While SVM algorithm takes a little time in comparison to Random Forest algorithm. SVM Radial Kernel is having second highest accuracy. So, we can say that Radial Kernel is best kernel in SVM four kernels, and Random Forest algorithm is best algorithm than other algorithms. Thus, interpretation of tweets indicates that Sh. Narendra Modi popularity is more in 2019 elections. Further, when we apply these algorithms on both politician dataset 2,500 then accuracy is higher than 7,500 dataset.

Table-4



Algorithm name	Accuracy on Narendra Modi dataset 2,500	Accuracy on Rahul Gandhi dataset 2,500	Accuracy on combined dataset 7,500
SVM-Kemel-Linear	71%	57%	63%
SVM-Kemel-Radial	72%	58%	64%
SVM-Kemel-polynomial	68%	59%	61%
SVM-Kemel-Sigmoid	68%	58%	60%
Random Forest	73%	58%	66%
Naïve Bayes	67%	51%	46%

In table-4, when we apply algorithms on same size of dataset Sh. Narendra Modi and Sh. Rahul Gandhi, the results accuracy is different. Like Sh. Narendra Modi dataset accuracy is 73% from Random Forest and Sh. Rahul Gandhi dataset accuracy is 59% from SVM Kernel Polynomial while when we apply algorithms on combine dataset which is higher in size from other dataset, the accuracy is 66% from Random Forest algorithm. The accuracy of 7,500 dataset is 66% which is lower in comparison to Sh. Narendra Modi dataset 2,500.

Table-4 has defined with accuracy on different datasets that produces two more conclusion; first, accuracy of algorithm does not only depends on size of dataset, it depends more on content of dataset, and second, it is not necessary that algorithms provide same result on every dataset which means that it is not necessary that Random Forest algorithm always produce high accuracy on each dataset. If dataset contents are changed like audio, video, or numerical based then may be the other algorithm perform the high accuracy. So, we can say that the algorithm accuracy also depends more on dataset contents. However, outcome stated in this work need extensive perfection based on large social media posts of tweeter as datasets using machine learning algorithms with clear-cut purpose.

VI. CONCLUSION AND FUTURE WORK

The aim is to analyze sentiments of public regarding Indian politicians using social media network as tweets. The purpose of this is to develop the best practice of analysis from available huge amount of text data on twitter as social media posts from 2014 to 2018 using supervised machine learning algorithms to predict popular leader of 2019 election. Results of three machine learning algorithms are compared, and Random Forest algorithm is found suitable with more accuracy among others. Thus, the content of dataset is also having impact on accuracy of algorithm. In future, more algorithms are used like Tree Based Algorithms, Maximum Entropy Algorithm and many more. More dataset from many other sites like Facebook, YouTube, Instagram etc are collected to establish the trends. This may help for better analysis of sentiments. However, outcome stated in this work need more extensive perfection based on large social media posts of tweeter as datasets using supervised machine learning algorithms with clear-cut purpose to ascertain to analysis of six sentiments.

REFERENCES

[1] Jaspreet Singh, Gurvinder Singh and Rajinder Singh "Optimization of sentiment analysis using machine learning classifiers", Human- centric computing and information sciences, Department of computer science, Guru Nanak Dev University, Amritsar India, 11 December 2017. https://hcis-journal.springeropen.com/track/pdf/10.1186/s13673-017-0116-3

[2] D. Krishna madhuri "A machine learning based framework for sentiment classification: Indian railway case study" Department of computer science and engineering, IJITEE, Vol-8, Issue 4, February 2019 Hyderabad, India.

[3] Erik boiy, marie-francine moens "A machine learning approach to sentiment analysis in multi-lingual web exts", department of computer science katholieke university Leuven, Belgium 13 august, 2008. https://www.researchgate.net/publication/220479915_A_Machine_Learning_Approach_to_Sentiment_Analysis_in_Multilingual_Web_Texts.

[4] Ali hasan, sana moin, ahmad karim, and shahabuddin shamshriband " Machine learning based sentiment analysis for twitter account", Department of computer science, information technology, management of science technology development, university of AIR Pakistan, bahauddin zakariya in Pakistan, ton duct hang in Vietnam, MDPI 27 February 2018. https://www.mdpi.com/2297-8747/23/1/11

[5] Malak Abdullah, Mirsad hadzikadic "Sentiment analysis of twitter data: Emotions revealed regarding Donald trump during the 2015-2016 primary debates" college of computing and informatics, university of North Carolina at charlotte North Carolina, IEEE 2017 International conference on tools with artificial intelligence, PP. 760-764, U.S. https://www.researchgate.net/publication/325635311_Sentiment_Analysis_of_Twitter_Data_Emotions_Revealed_Regarding_Donald_Trump_during_the_2015-16_Primary_Debates.

[6] Abhishek Kaushik and Sudhanshu Naithani "A comprehensive study of text mining approach" Kiel university of applied science, kurukshetra university, IJCSNS, Vol-16 No.2, February 2016, PP.69-76.

[7] R Documentation -https://www.rdocumentation.org

[8] Vimalkumar B. vaghela, Bhumika M. jadav "Analysis of various sentiment classification techniques" L. D. College of engineering Ahmadabad, India.

[9] Abdullah Alsaeedi, Mohammad Zubair Khan "A Study on Sentiment Analysis Techniques of Twitter Data" college of computer science and engineering Taibah University Madinah, KSA, IJACSA, Vol-10, No. 2, 2019, PP. 361-374.

[10] Parul Sharma, teng-sheng moh "Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter" Department of computer science, San Jose state university, IEEE International conference on big data 2016, PP. 1966-1971, CA, USA.

[11] https://data-flair.training/blogs/svm-kernel-functions

[12] https://en.wikipedia.org/wiki/Kernel_method

[13] https://blog.easysol.net/machine-learning-algorithms-4

[14] Gerard Biau , "Analysis of a Random Forests Model" , Universite Pierre et Marie Curie – Paris VI ´ Boˆıte 158, Tour 15-25, 2eme ` etage ´ 4 place Jussieu, 75252 Paris Cedex 05, France. https://www.yumpu.com/en/document/view/35311802/analysis-of-a-random-forests-model-universite-pierre-et-marie-

[15] https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm.

[16] Niharica choubey, Vilender Kumar, Sanjay Kumar Gupta, Use of Tidytext Lexicons Approaches for Sentiment Analysis, IARJSET, Vol. 7, Issue 1, January 2020, PP. 58-64.