Sentiment Analysis Using Naïve Bayes Classifier

Kavya Suppala, Narasinga Rao

Abstract: Twitteris a web service and social communication platform which allow users to address their tweets in different domains. Public can easily and efficiently explicit their perspectives and ideas on a wide variety of cluster on topics via social networking websites. As online data is abundant through different platforms like social networks, twitter, Facebook, etc... Analysing the data is of paramount importance in drawing inference from the data. Hence, in our research, we try to perform sentiment analysis on twitter data by using a Naive Bayesian algorithm. By using our model, we can measure the customers opinions and perceptions and can be enhanced to any desired level depending on the data gathered from on line resources.

Key words: twitter data, machine learning, and naïveBayes classifier.

I. INTRODUCTION

The history of the Internet has changed the way people opine on their perceptions. Now a days it is all done through different blog posts, online discussion forums, product review websites etc. People usually depend on user-generated content on any product to a great extent when it comes to perform any desired action. When people want to buy a product through online, they will first look up its reviews in that particular product website through online, before making up a decision. Some analysis is to be done on all these reviews so that the final outcome says whether the product is good to buy or not. There are different sentiment analysis techniques that are available with many applications for different domains, like in business to get a feedback for products from customers. Knowledgebase and Machine learning techniques are two techniques that are mainly used for sentiment analysis. In the case of Knowledgebase approach this requires a large database with predefined emotions and an efficient and effective knowledge representation for identifying sentiments. In the case of Machine learning approachdoesn’t require any predefined set of emotions, this makes use of a training set in order to develop a sentiment classifier which classifies sentiments from the tweets and so machine learning approach is rather simpler than knowledgebase approach. Actually, there are different machine learning techniques that are used to classify data i.e., they are naïve Bayes classifier, support vector machine, decision tree, random forest, neural networks etc.

Classification is a technique which is used to perform classification on different sets of data into different classes. These classification techniques are divided into two categories Supervised and unsupervised. In supervised learning approach, there will be a teacher which makes the computer to learn from the labelled input data that is given to it and then makes the computer to use this learning which is used to classify output data. In this the dataset maybe in the form of bi-class i.e., like identifying whether the statement is positive or negative or it may be multi-class too i.e., statement may be positive or negative or neutral. In unsupervised learning approach, computer learns with the unlabelled input data and which is used in grouping the data for example in cluster analysis.

Many papers have discussed the models which have been developed to perform analysis on Twitter data. Gaurav D Rajurkar, Rajeshwari M Goudar [11] they have developed a model for Twitter sentiment analysis using HADOOP which reduces the time delay and cost on using a large amount of data by using an efficient Apache Open Source Product for Twitter data analysis with no extra work like scraping, cleansing and also provides the speedy data downloading approach for analysis.

EfthymiosKoulompis, Theresa Wilson, Johanna Moore [13] have collected in total 3 sets of datasets i.e.,Hash tagged (HASH) dataset compiled from Edinburgh twitter corpus for development and training, 2nd is emoticon dataset from http://twittersentiment.appspot.com and the last one is manually annotated dataset for evaluation produced by an iSieve corporation. They have used different features namely n-gram, lexical, parts-of-speech etc. Their main aim is to investigate the utility linguistic features and its usefulness for detecting the sentiment on twitter messages. Finally, results show that the parts-of-speech feature was not useful in sentiment analysis in this microblogging domain. Saif, Hassan; He, Yulan and Alani, Harith [12] a method was developed that can automatically collect the corpus and train the sentiment classifier based on multinomial naïve Bayes uses 2 features i.e., n-gram and POS-tagsand the classifier determines the classes of each tweet. Their main aim is to prove that a developed technique is more efficient than previously proposed methods.

II. LITERATURE SURVEY

HumaParveen and Prof.ShikhaPandey [1] did sentiment analysis on movie dataset by downloading the tweets on developing a Twitter API.They have used the Hadoop framework for processing the dataset. Classification is done using the Naïve Bayes algorithm and its performance is increased by pre-processing the tweets. The final results shows the classification of text in their required classes

Revised Manuscript Received on June 05, 2019

Kavya Suppala, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.

Narasinga Rao, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.

with accurate performance.

Neethu M S and Rajasree R [2] performs analysis on tweets based on some specific domain using different machine learning techniques.They tried to focus on problems that are faced during the identification of emotional keywords from multiple keywords and difficulty in handling misspellings and slang words. So a feature vector is created whose accuracy is tested using naïve bayes, SVM, maximum entropy and ensemble classifiers.

Bac Le and Huy Nguyen [3] built a model to analyze the sentiment on Twitter using machine learning techniques by applying effective feature set and enhances the accuracy i.e., bigram,unigram and object-oriented features. The classification of tweets is done using 2 algorithms i.e., Naïve Bayes classifier and Support vector machines(SVM) whose accuracies are tested by calculating precision, recall and f-score and also shows same accuracy.

Sayali P. Nazare, Prasad S. Nar, Akshay S. Phate, Prof.Dr. D. R. Ingle [4] have created a dataset by twitter API and collected all tweets regarding the topic blue whale game. Their main aim is to perform analysis on sentimental tweets. They have used Naïve Bayes,Support vector machines,Maximum entropy and Ensemble classifier.SVM and Naive Bayes classifiers are implemented using MATLAB built-in functions. Maximum Entropy classifier is implemented using MaxEntsoftware. Based on comparative results Naïve Bayes has better precision and slightly lower recall and accuracy i.e., 89% and other classifiers are having similar accuracy levels i.e., 90%. The result shows the pie-chart which is representing the positive,negative and neutral hashtags with percentages.

Dey, Lopamudra&Chakraborty [14] have collected 2 sets of dataset they are movie reviews and hotel reviews by using 2 classifiers naïve Bayes and K-NN. Their aim is to check which classifier gives better results on both datasets. The experimental results shows that the naïve Bayes classifier gives better performance in the case of movie reviews dataset and on considering hotel reviews dataset both classifiers shows approximate results. Finally, naïve Bayes classifier is better for movie reviews classification.

R. Dey and S. Chakraborty [15] developed a new technique which predicts the weather conditions from air polluted dataset. Then applied convex-hull technique suitable for dynamic databases where the climate data are changed frequently. The incremental DBSCAN clustering is used which performs clustering of new data that is inserted and a protocol is used to give the weather prediction. The results gives the accuracy of the model based on hit and miss.

Anuja P Jain and Asst. Prof Padma Dandannavar [5] proposed a method for performing sentiment analysis on dataset from Apple, Bank and BSNL of size ranging from 200-4000. In which 3/4th of dataset is taken for training and the remaining 1/4th is taken for testing. They have used 2 different classification techniques i.e., Multinomial Naïve Bayes classifier and Decision tree classifier. Pre-processing of tweets is done using the feature extraction. They have used Apache Spark framework which is scalable and gives faster accurate results. Finally, the Decision tree classifier gives 100% accuracy, precision,recall and f1score.

Mejova, Yelena [6] discusses about the overview of the sentiment analysis i.e., they give brief description about the word sentiment and novelty methods that are used to perform analysis on emoticoncharacter text by covering all the challenges that are faced during analysis of product or text etc.Leena A Deshpande, M.R. Narasinga Rao [17] developed a method that determines the variance in the data and retrain the model confer to the drift that is identified. They have used 2 techniques weight-based features and n-Gram that identifies the unnamed labels in the text which improves its accuracy.

T. Sajana, M.R.Narasingarao[18] they have performed survey on detection and prediction of malaria disease using various machine learning techniques, Image Processing techniques.They have observed that machine learning techniques are mostly applicable for critical diagnosis of malaria.

Boiy, Erik, Hens, Pieter, Deschacht, Koen &Moens, Marie-Francine, Marie-Francine [7] show cased the usefulness of classification and also the overview of some methods by calculating the accuracy of each method shows encouraging results. Indicates the challenges that are faced while gathering the data from World Wide Web. Niu, Zhen, Zelong Yin, and Xiangyu Kong [8] performs classification of text having sentiments using web services. They have developed three algorithms but improved only one to enhance the overall efficiency. Also introduced new method based on naviebayesclassifier which classifies the text in microblog with various sentiments having highest efficiency. F. Neri, C. Aliprandi, F. Capeci, M. Cuadros and T [10] shows the sentiment analysis study of 1000 facebook posts on newscasts by knowledge mining system and considering the data from Auditel regarding newscast audience, correlation analysis of Facebook with measurable data which is available in public.

III. NAÏVE BAYES CLASSIFIER

Algorithm:

1. Consider a training data set D consists of documents which belongs to different classes say class A and B.
2. Prior probability of both classes A and B is calculated as shown
   Class A=number of objects of class A / total number of objects.
   Class B=number of objects of class B / total number of objects.
3. Now calculate the total number of word frequencies of both classes A and B i.e., n_i
   n_a = the total number of word frequency of class A.
   n_b =the total number of word

frequency of class B.

4. Calculate the conditional probability of keyword occurrence for given class
   P(word1 / class A) = wordcount / n_i(A)
   P(word1 / class B) = wordcount / n_i(B)
   P(word2 / class A) = wordcount / n_i(A)
   P(word2 / class B) = wordcount / n_i (B)
   ………………………………………
   ………………………………………
   P(wordn / class B) = wordcount / n_i (B)
5. Uniform distributions are to be performed in order to avoid zero frequency problem.
6. Now a new document M is classified based on calculating the probability for both classes A and B P (M/W).
   a) Find P(A / W) = P(A) * P(word1/class A)* P(word2/
      class A)……* P(wordn / class A).
   b) Find P(B / W) = P(B) * P(word1/class B)*P(word2/
      class B)……* P(wordn / class B).
7. After calculating probability for both classes A and B the class with higher probability is the one the new document M assigned.

IV. PROPOSED WORK

In proposed work, we have discussedhow a sentiment is extracted from a tweet/text using Twitter dataset. It is a place where the users posts their views and opinions based on the situation. The main objective of our proposed system is to perform analysis on tweets having sentiment which causes the great help to business intelligence on predicting the future**.**This paper addresses the sentiment analysis on twitter dataset; that is at first classification is performed on tweets using naïve bayes classifier. Each tweet is represented in the form of sentiment asserted in terms of positive, negative and neutral. Performing sentiment analysis is vital which is used to find out the pros and cons of their products in the market by public that results in improving their business productivity. The aim of this project is to develop a classification technique using machine learningwhich gives accurate results and automatic sentiment classification of an unknown tweet by predicting the future. In this paper, sentiment analysis is done on Twitter data. The dataset is collected which contains 65536 tweets these tweets are collected based on the situation on all topics. There are different attributes in the database such as item-id, sentiment, sentiment source, sentiment text but sentiment text has been considered for our proposed research. The first attribute item-id contains the id of the tweet, the second attribute sentiment represents the Boolean value (1 or 0) i.e., the tweet containing sentiment is taken as 0 and tweet without any sentiment is declared as 1, and the third attribute sentiment source represents the source from the tweet is taken and of maximum length 140 characters, and the last attribute sentiment text represents the text or tweet based on all situations either containing sentiment or not. Our main aim is to perform analysis on these tweets and conclude the tweets which are positive and negative.

So in order to classify data first, we need to perform the following steps.

- Tokenization: It is a method that divides the variety of document into small parts called tokens. These tokens may be in the form of words or numbers or punctuation marks.
  Ex: it is going to rain today
  After performing tokenization the sentence is divided into tokens as follows
  "It", "is", "going", "to", "rain", "today".

- Stop words: These are the common words that are to be ignored which reduces the size of the dataset also the no of words (tokens). In our programming language pythonwe use a tool called natural language tool kit(NLTK) in which there is list of stop words in 16 different languages.
  Ex: I like dancing, so I dance.
  After removing stop words the sentence will be as follows
  Like,dancing,dance.

- Bag of words concept is applied to these tokens.

- Finally, our classification technique Naïve Bayesian classifier is applied which calculates the probability of all words in the document and gives the result i.e., probability of each tweet in both positive and negative.
- Results show the probability of each tweet saying whether the tweet is either positive or negative.

A. Bag-of-words

A bag-of-words is a representation of text that describes the occurrence of words within a document. The occurrence of words is represented in a numerical feature. It is a way of extracting features from the text for use in modelling, such as with machine learning algorithms.The approach is very simple and flexible and can be used for extracting features from documents.But there is some complexity on twocases i.e., one is on designing the vocabulary of known words and the other is on scoring the presence of known words.

Let us consider there are 2 classes i.e., positive class and negative class. Each class contains some words that is positive class contains some bag of positive words (slow, fine, good, fantastic) and negative class contains some bag of negative words (hate, terrible, heavy). We will give the input as a text/sentence and starts counting the frequency of each word in the document and this gives the result whether the text/sentence belongs to positive class or negative class.

Example: "it is going to rain today"

“today I am not going outside"

“I am going to watch the season premiere”

Now tokenization is performed on these lines then we get

Line 1:

It, is,so, hot, today.

Line 2:

Today, I, am, not, going, outside.

Line 3:

I, am, going, to, watch, the, season, premiere.

Now we have to maintain a table which contains 2 columns with attributes word and counts i.e., we are going to count frequency of each word in the document.

Table1: Frequency of each word represented as count.

Word	Count
It	1
Is	1
so	1
hot	2
Today	1
I	2
Am	2
Not	1
outside	1
Watch	1
The	1
Season	1
Premiere	1

Now after completion of the table, we came to know the words that are frequently occurred and the words with rare cases. So due to this, we don't really want to consider all the different words that appear in different documents. We consider only some fixed words. As from the above table, there is a total of 14 words but after that, there are only 10 most frequent words which reduce the comparisons.

B. Application of sentiment analysis

Naïve Bayes classifier is one of the supervised classification technique which classifies the text/sentence that belongs to particular class. It is the probabilistic algorithm which calculates the probability of each word in the text/sentence and the word with highest probability is considered as output.

- Let us consider a document a
- A document a with a set of classes B = { b_1, b_2, … , b_n }
- Consider a training set having m documents which is pre-determined that belongs to a particular class.

Now we train our classification algorithm using this training set and we get trained classifier. By using this trained classifier we can classify the new document.

C. Bayesian Theoremapplied to Documents.

For a document **a** and a class **b** using Bayesian theorem,

P (b | a) = [p (a | b) * p (b)] / [p (a)]

- The term p (a|b) is represented as

Now representing the document **a** as a set of features (words or tokens) **x_1, x_2, x_3 …**
We can then re-write **P (a | b)** as:
P(x1, x_2, x_3… x_n | b)

- P (b) is defined as **total probability** of a class. Which gives the frequency of class b

Example:let us consider two classes **positive and negative** without analysing the input document the probability of text/sentence is calculated which results whether the text is positive or negative

The calculation is done by counting the relative frequencies of each class in a corpus.

E.g. out of 10 reviews we have seen, 4 have been classified as **positive**.

P (positive) = 4 / 10

Example: considera tweet or a sentence such as “It is going to rain today" now. We are going to apply naïve Bayes classifier and say whether the sentence is either positive or negative.

By considering the bag of words concept i.e., which contains some positive and negative words and their frequency counts. Now comparison is performedamong each word in the sentence and positive and negative words in bag of words. Probability is calculated on both positive and negative words, the words which is having highest probability is taken into consideration i.e., if the positive words having highest probability, then the tweet is considered as positive and vice versa.

P(it is so hot today) = P(it)*P(is)*P(so)*P(hot) *P(today)

By removing stop words, the words will be reduced.

P (it | positive)*P(is | positive)*P (so | positive)*P (hot | positive)*P(today | positive)

After calculating the probability of the above statement we get some overall probability for positive words

P (it | negative)*P(is | negative)*P (so | negative)*P (hot | negative)*P(today | negative)

After this, we will get some overall probability for negative words.

Now by comparing both probabilities the words having the highest range is taken into consideration. By this, we can say whether the new input tweet or sentence is positive or negative.

IV. RESULTS AND ANALYSIS

Bar charts:

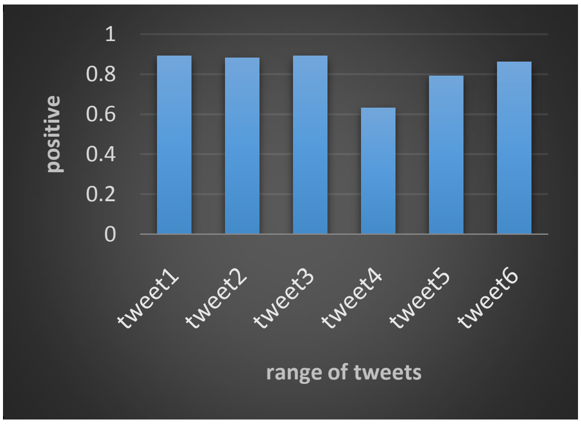

Figure 1: Graph showing the relation between range of tweets and its % of positiveness

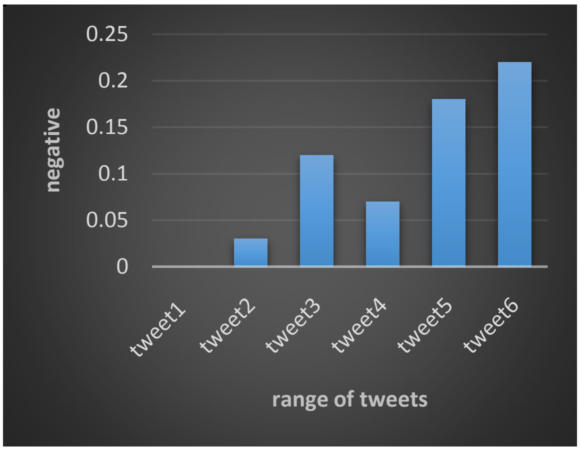

Figure 2: Graph showing the relation between range of tweets and its % of Negativeness

Table 2: Probabilities of each tweet in both positive and negative.

Tweets	positive	Negative
i think i lost my spanish book!!	0.7	0.1
life can be so complicated. i know god has a plan .... i just have to trust in him	0.6	0.2
i don't know what to do..	0.8	0.1
.i did not get cast in but i will keep looking for some dance opportunities. i need some exercise.	0.7	0.1
the ground around me is wet... i looked up to see if any drops fall on my face...it was drizzle all around but i am missing my drops	0.4	0.1
i forgot to bring my new sharpie pens to school today..booo!	0.5	0.08
:'(this is so horrible	0.6	0.3
:'(where ius that one person that can make me smile	0.4	0.3
@alittlebit - temperature is down, but the pain is still bad. hoping antibiotics kick in soon! thanks for asking	0.6	0.2
@anyamanda said i tweets less these days so, hi!	0.6	0.1
@benboychuki hope it's not actually on here birthday that she's getting shots.	0.5	0.4
@billbeckett's signature has 100% rubbed off my mondayeyes bracelet. at least i have plenty of autographs from him, but still. sad day	0.3	0.4
@blondediva11 what happened while i was gone? anybody get arrested http://tr.im/kofz	0.3	0.4
@breegeeki'll know if u can't find a replacement 4me. i'll still pay u if not. i won't let u hang	0.6	0.2

@canuckgrrrl seeing a guy dressed up as a banana totally made my day. phallic symbol ftwhttp://tr.im/mvtm	0.3	0.3
@carole29 yep coz of the england game on Wednesday	0.4	0.1
youaintgotta apologize. i know i'm not ugly.	0.5	0.0
yep,i have things my way all the time. i am selfish. everything revolves around me. yep. im a bitch. great for me	0.2	0.5
.i don't wanna get rid of twitter you guys are too cool...	0.6	0.2
i try... i try so hard... and i seem to get no credit for it.... wats the prob?	0.5	0.35
wide awake, but dont actually wanna go to school	0.6	0.2
imgivin myself 5 minutes to turn that frown upside down	0.6	0.0
icant even say a word right now	0.4	0.4
why am i so sad??...becuz of u or becuz of the exam??.....i think..becuz of u.	0.5	0.2
just found out that a friend from my childhood has passed away... may he rest in peace	0.3	0.3
sniffles... since idont get to leave this fucking house...aw great the power's gone. andits getting dark.. puurfect!	0.5	0.3

V.CONCLUSION

In conclusion, we have developed a model which performs sentiment analysis on Twitter data using Machine Learning Technique. The model that was proposed in this research built by using Natural Language Tool Kit (NLTK) on the dataset containing tweets. Bag of words concept is used which contains both positive and negative words separately. The classification was done using Naïve Bayes classifier by calculating the probability of new input data and the tweet with the highest value is considered as either positive or negative. However, we chose an effective twitter feature dataset which enhances the effectiveness and accuracy of the classifier. This model can further enhanced to any desired level if one wants to by incorporating more features in the database.

REFERENCES

1. HumaParveen and ShikhaPandey*“Sentiment analysis on Twitter Data-set using Naive Bayes algorithm*” 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) page 416-419 @article{Parveen2016SentimentAO}.
2. M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, Tiruchengode, 2013, pp. 1-5.
3. Le B., Nguyen H. (2015) “Twitter Sentiment Analysis Using Machine Learning Techniques”. In: Le Thi H., Nguyen N., Do T. (eds) Advanced Computational Methods for Knowledge Engineering. Advances in Intelligent Systems and Computing, vol 358. Springer, Cham
4. Sayali P. Nazare, Prasad S. Nar, Akshay S. Phate, Prof.Dr. D. R. Ingle*“Sentiment Analysis in Twitter”*International Research Journal of Engineering and Technology (IRJET)Volume: 05, Jan-2018.
5. Anuja Prakash Jain and Padma Dandannavar “Application of machine learning techniques to sentiment analysis” 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology(iCATccT) pages 628-632 article{Jain2016ApplicationOM}
6. Mejova, Yelena. (2019) “Sentiment Analysis: An Overview”.
7. Boiy, Erik, Hens, Pieter, Deschacht, Koen &Moens, Marie-Francine, Marie-Francine. (2007). “Automatic Sentiment Analysis in Online Text”. ELPUB2007. Openness in Digital Publishing: Awareness, Discovery, and Access - Proceedings of the 11th International Conference on Electronic Publishing held in Vienna, Austria 13-15 June 2007 / Edited by Leslie Chan and Bob Martens. ISBN 978-3-85437-292-9, 2007, pp. 349-360.
8. Niu, Zhen, Zelong Yin, and Xiangyu Kong. "Sentiment classification for microblog by machine learning." In 2012 Fourth International Conference on Computational and Information Sciences, pp. 286-289. Ieee, 2012.
9. J. Ren, S. D. Lee, X. Chen, B. Kao, R. Cheng, and D. Cheung, "Naive Bayes Classification of Uncertain Data," 2009 Ninth IEEE International Conference on Data Mining, Miami, FL, 2009, pp. 944-949.
10. F. Neri, C. Aliprandi, F. Capeci, M. Cuadros and T. By, "Sentiment Analysis on Social Media," 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, 2012, pp. 919-926.
11. Gaurav D Rajurkar, Rajeshwari M Goudar, "A speedy data uploading approach for Twitter Trend And Sentiment Analysis using HADOOP", HADOOP, 2015 International Conference on Computing Communication Control and Automation. Pages 580-584.
12. Saif, Hassan; He, Yulan and Alani, Harith(2012),"Semantic sentiment analysis of Twitter," in The 11thInternational Semantic Web Conference (ISWC 2012),11-15 November 2012, Boston, MA, USA. Pages 1320-1326.
13. EfthymiosKoulompis, TheresaWilson, Johanna Moore (2011),*“Twitter Sentiment Analysis: The Good the Bad and the OMG!,"* in The Fifth International AAAIConference on Weblogs and Social Media. Pages 538-541.
14. Dey, Lopamudra&Chakraborty, Sanjay & Biswas, Anuraag& Bose, Beepa& Tiwari, Sweta. (2016). “Sentiment Analysis of Review Datasets Using Naïve Bayes and K-NN Classifier”. International Journal of Information Engineering and Electronic Business. 8. 54-62. 10.5815/ijieeb.2016.04.07.
15. R. Dey and S. Chakraborty, “Convex-hull &DBSCAN clustering to predict future weather”, 6th International IEEE Conference and Workshop on Computing and Communication, Canada, 2015, pp.1-8.
16. Sathyadevan, Shiju& S, Devan & S Gangadharan, Surya. (2014). “Crime Analysis and Prediction Using Data Mining”. 10.1109/CNSC.2014.6906719.
17. Leena A. Deshpande, M.R. Narasingarao “Addressing social popularity in twitter data using drift detection technique” Journal of Engineering Science and Technology Vol. 14, No. 2 (2019) 922 – 934.
18. Tiruveedhula, Sajana &ramanarasingarao, Manda. (2017). “Machine learning techniques for malaria disease diagnosis - A review” Journal of Advanced Research in Dynamical and Control Systems. 9. 349-369.