Turkish Journal of Computer and Mathematics Education Vol.12 No. 7 (2021), 1669-1675

Research Article

L2b: Lexicon Boosted Bayesian Classification For Popularity Prediction Of Movies With Improved Accuracy Using Twitter Corpus

[a]Gunasekaran, [b]Prakash V S, [c]Amitabh Wahi, [d]N Mohammed Imtiaz

[a]Professor, Department of Computer Science and Engineering
Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences
Chennai – 602 105, Tamilnadu, India
Email: gunasekaranm.sse@saveetha.com
[b]Assistant Professor, Department of Computer Science, Kristu Jayanti College
Bengaluru – 560077, Karnataka, India
Email: vsprakash@kristujayanti.com
[c]Dean, Research and Development Cell
Guru Nanak Institutions, Khanapur (V), R R District, Hyderabad, India
Email: deanrnd@gnuindia.org
[d]Assistant Professor, Department of Computer Science and Engineering
GRT Institute of Engineering and Technology, Tiruttani – 631209
Email: imtiaz4687@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract- Social networking plays a vital role in the research phenomenon. It helps in acquiring the most recent news and information consistently with settling on ground-breaking choices and expectations. There are enormous expressions of people opinion in addition to what's circumventing the world. Twitter, is a well known blogging site where individuals place their perspectives and inclinations based on their interest. This paper presents a Lexicon Boosted Bayesian (L2B) classification for popularity prediction of movies based on twitter corpus. The proposed approach predicts the accomplishment of the film utilizing perspective individuals that might actually be accomplished by assessment examination with Naïve Bayes classification and Lexicon approach. In addition, it focuses with the representation of information in R language which stands the best in speaking to the investigation in pictorial arrangement and a guide perception that depicts the area of the tweets where it has come from. The result analysis proves the effectiveness of proposed technique in comparison with the existing approaches for predicting the success of the movies.

Keywords: Text mining; sentiment analysis; twitter data; predictive analytics; Lexicon; Naive Bayes

1. Introduction

Gone are the days when people had to watch movies straight away at the theatres without knowing how the movie has come up. Then the days and even today people have been listening/seeing reviews after reviews of same movie using various social media such as radio, newspaper, television, magazines and social networking sites to arrive at an inference whether the movie is one worth watching at theatres Dubey PK and Agrawal S (2013), Jain V (2013), Sitaram et al. (2010). Now they don't even believe these media because they knew media usually bragabouta movie even if it has not come up well. It is not necessary to say if it is a big star's movie in the industry. The world which is getting even busier has no time now to do this and it will be obsolete in months from now. So, the future is this technology that would predict the success of a movie in one click based on tweets of people on twitter. Because public are the ones for whom the movie is made for. Hence, obtaining views of people about a movie directly through twitter helps come up with a better prediction Mohammed M F et al. (2018), Ravi K and Ravi V (2015).

The Bayesian Classification Liangxiao J (2009) is a sophisticated, statistical and suitable method for classifying large data sets. It is a primary probabilistic approach and permits us to define uncertainty through a direct method by determining the probabilities of the result. This approach can resolve diagnostic and predictive problems. In addition, this model supports with an useful perspective for understanding and evaluation of many existing machine learning techniques Domingos P (1997). It determines unambiguous probabilities for hypothesis and also it is strong and robust against to noise in input data. This classification model has been used as a probabilistic learning method and also the survey ensures that this is the most successful known algorithms to categorize text documents based on survey. Lexicon a piece of instant message is addressed as a group of words. Following this portrayal of the message, slant esteems from the word reference are allocated to all sure and negative words or expressions inside the message.

Nowadays, lot of research is going on for prediction of movies. All the prediction techniques

concentrate on the users support and their interest on various movies, whereas few of the movies use digital media for prediction. However, minimal work has been done on using lexicon based sentiment analysis. The volume of information that is available about the movies in the web makes its important aspirant for forecasting, classification, and knowledge discovery and also machine learning and deep learning. There are many research works are being done on movies dataset related to their popularity, reason for popularity, reviews, and its ratings over the web. Prediction of movie popularity is of high significance to film industry and movie producers. But still they are not sure about whether there movie will be successful or not; when to release the film and how to promote it Darin I and Minh T N (2011), Nithin VR (2014). In this paper, an efficient technique for popularity prediction of movies based-on twitter corpus has proposed that adopts supervised learning such as Naïve Bayes classification technique and lexicon-based sentiment analysis for the prediction of movies.

Below are the factors that state why the proposed work is considered to be important:

(i) Raising a pre-release hype among people by such effective visualization turns out to be a success factor both fame-wise and monetary-wise for actors, multiplex owners and distributors.
(ii) This work will help people to spend a quality time of 3 hours at theatres.
(iii) Release of even a single movie at an appropriate time could even bring in a huge difference of corers of rupees of profit.
(iv) Unlike Hollywood, much research has not been done for regional movies in India.

This paper has been structured as: the section 2 discusses the existing machine learning algorithms and techniques that talks about movie classification to predict the popularity based on various metrics, methods and other constraints. The section 3 describes the proposed methodology for the movie prediction and the reason for their popularity. The section 4 depicts the results and discussions and performance analysis of the proposed system. Finally, section 5 concludes the paper.

2. Related Works

Aryan et al. (2020) proposes a text mining approach that orders the gatherings into different groups, with the center point of each addressing an area of interest discussion inside the current period of time by consolidating K-means clustering and support vector machine. The authors utilize a scope of 31 diverse theme gatherings and 220,053 posts. The proposed model got exceptionally reliable outcome. Nirmala Devi and Murali Bhaskaran (2012) proposes an area of interest foreseeing approach, in this model the promoting division comprehends what their particular clients' opportune concerns with respect to the products and ventures. Accordingly the productive discovery of area of interest gatherings dependent on estimation examination may make web informal organization individuals advantage in the dynamic interaction. Sameer T and Tushar P (2014) propose a Hype investigation based methodology that shows how web-based media could be used to foresee future results. The proposed model uses the rater of jabber from tweets from the populace site twitter and developed numerous direct relapse models for anticipating film industry incomes of motion pictures ahead of their delivery. This work likewise shows how friendly average communicates an aggregate insight which, when appropriately tapped, can yield a very amazing and exact pointer of future results.

Dipik DG and Bijith M (2015) propose sentiment analysis and prediction algorithms to dissect the presentation of Indian motion pictures dependent on information got from web-based media locales. The creators utilized Twitter4j Java API for separating the tweets through verifying association with Twitter sites and put away the extricated information in MySQL data set and utilized the information for assessment investigation. Hoda S et al. (2013) have done a big data analytics utilizing constant tweets about motion pictures to gauge the income and accomplishment of films. The authors utilize an enormous dataset of tweets around 50 films and receive assessment investigation to inspect the clients' inclinations towards the motion pictures and exhibit that the film income forecast model is legitimate. Flex MSW, Soumya S and Mung C (2012) examines about the extraordinary admittance to general assessment on decisions, news, motion pictures, and so on In this investigation, the issue with regards to film audits on and contrasts the assessment of Twitter clients and that of IMDb and Rotten Tomatoes. This paper shows that Twitter clients are more supportive in their reviews across most motion pictures in contrast with other rating destinations.

3. Proposed Methodology: Lexicon Boosted Bayesian Classification

The proposed methodology is implemented based on Lexicon and Bayesian classification approaches. Fig. 1 show the L2B, during the first step lexicon has to be created with assigned polarity scores. Then these lexicons are to be grouped that becomes a large dataset that needs to analyzed for sentiment. During preprocessing or cleaning, identify the syntactic construction which is not useful needs to be removed. In Addition, stop-words and negations are removed. In feature generation and lexicon generation, the patterns that are most useful to capture the data are identified; this improves the quality and accuracy of the movie classification. During feature selection, select the features that are relevant and have the greater predictive influence. In addition, the important emotions are identified and incorporated into the lexicon dataset in order to improve the accuracy of

popularity prediction analysis. In training and validation, the most relevant tweets are identified from training dataset and represented in terms of attributes. For these attributes the classes (positive, neutral, negative) are checked and ensures.

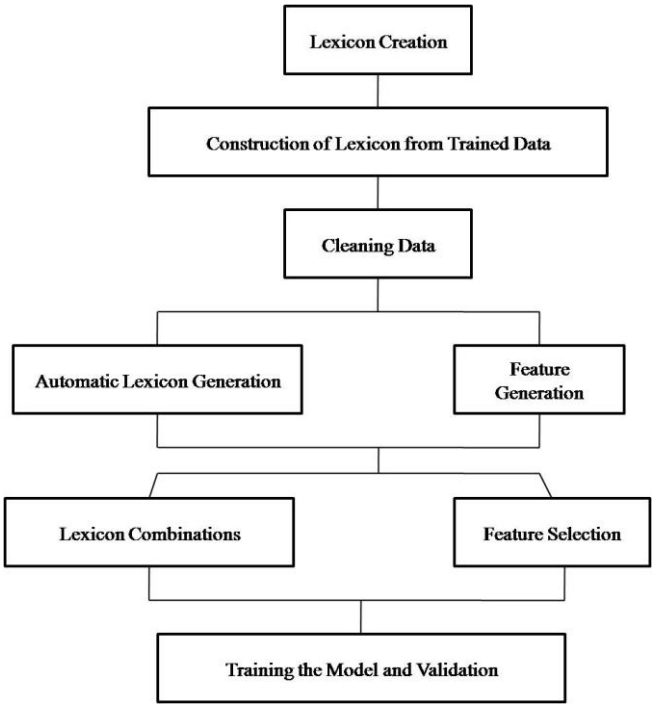

Fig. 1: Flowchart of L2B

3.1Lexicon Approach for Sentiment Analysis

Vocabulary is a book containing an in sequential order course of action of the words in a language and their definitions. Another way to deal with perform notion examination depends on a component of assessment words in setting. Assessment words will be words that are generally used to communicate positive or negative opinions, e.g., "great" and "awful". The methodology for the most part utilizes a word reference of assessment words to recognize and decide slant direction (positive, negative or unbiased). The word reference is known as the assessment vocabulary. The methodology of utilizing assessment words (the vocabulary) to decide assessment directions is known as the dictionary based way to deal with conclusion investigation. This methodology is proficient and can be utilized to break down text at the record, sentence or substance level. It is consequently material to our assignment too. In any case, Twitter information has built up its own attributes. Some of them are inconvenient to the vocabulary based methodology. For instance, emojis, casual articulations, shortenings, and so on are every now and again utilized in tweets. These articulations may have semantic/estimation direction yet they don't exist in an overall assessment dictionary. Allow us to see a tweet model, "I purchased iPad yesterday, simply lovvee it :-)". It plainly communicates a positive assessment on iPad by "lovvee" and the emoji ":-)".

In any case, the dictionary based technique would view the tweet as communicating no/unbiased assessment on iPad, since there is certifiably not an overall assessment word in the tweet. This prompts the low review issue for the vocabulary based strategy, which relies totally upon the presence of assessment words to decide the assumption direction. Albeit one may say that these extra articulations can be added to the assessment vocabulary, such articulations change continually and new ones are additionally showing up constantly following the patterns and designs on the Internet. Also, their polarities can be space subordinate. These issues make it hard to physically add them to the assessment dictionary. Without a complete dictionary, the supposition investigation results will endure. Accordingly, vocabulary based methodology or a word reference based methodology is remembered for this paper to have a comprehension of what it is and to execute it. The Fig.2 depicts the block diagram of lexicon approach for sentiment analysis.

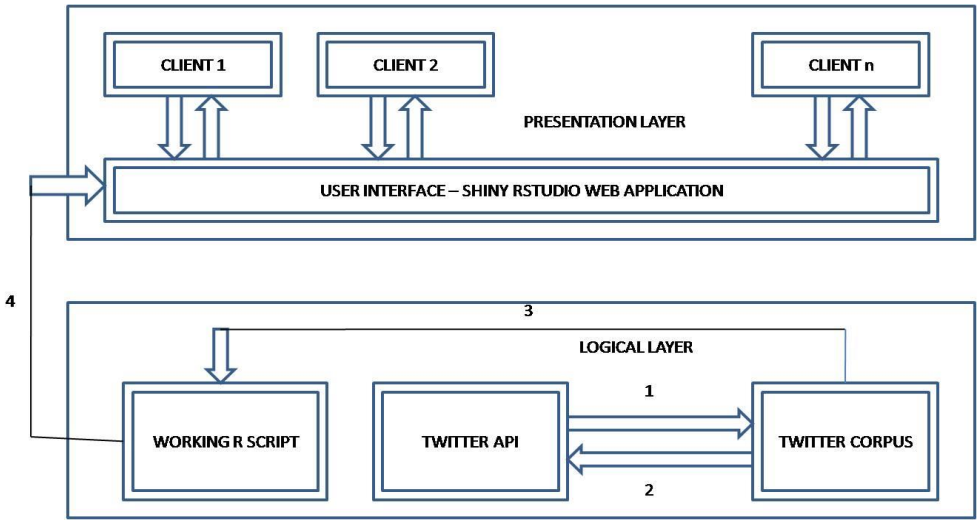

Fig. 2 Block diagram of Lexicon approach

3.2 Bayesian Classification

The proposed work helps in analyzing the state of mind of tweeple and achieves highly consistent results by applying Naive Bayes Model.

3.2.1 Creating a Twitter API at https://apps.twitter.com/

There two micro-blogging service of Twitter has been used that includes two RESTful APIs such as TEST API and Search API. The first one permits developers to access large Twitter dataset which includes update timelines, status data, and user information. The second one supports developers to connect with Twitter search and trends data. Hence, in order to extract tweets, we need a Twitter account and hence a Twitter API to get our own access keys and authentication to use twitter data in our application.

3.2.2 Establishing Connection to Fetch Tweets from Twitter

This establishes an formal connection with Twitter for getting tweets based on various search parameters. It requires installation of R libraries such as "ROAuth" – an open standard authentication for R interface and "httr" – to handle http requests.

3.2.3 Retrieval of Tweets to R Script

Once the connection is successfully established, we can fetch tweets with the help of twitter API to the working R script. One can extract a maximum of 3200 tweets at a time from the site. If required more, execution of set of code (that does fetching) can be done in a loop.

3.2.4 Sentiment Classification of Tweets

We can now analyze the sentiment of tweets using Naive Bayes algorithm, a supervised machine learning text classification algorithm that comes as a built-in algorithm in R language. All we have to do is to just use the sentiment" library in R, specifying the algorithm as "bayes" with required prior probability (that ranges between 0.0 – 1.0) and giving the tweet texts as input. With this library, we can make use of two functions "classify_emotion()" that has 6 built-in emotions such as joy, anger, sad, disgust, fear, surprise with which we can describe the state of mind of the tweeple and "classify_polarity()" that classifies polarities as positive, negative and neutral with which we can visualize the critics about the movie. Finally, we can visualize the emotions in a bar chart with x label as 'emotions' and y label as 'number of tweets with the respective sentiment' and visualizing the polarity in a pie chart showing the volume of critics in the tweets which is done with the help of "rCharts" library – an interactive Javascript data visualization library in R.

3.2.5 Formation of Word Cloud Based on the Occurrences of the Word in the Tweets

It is a graphical representation of repeatedly used words in a group of text files. The height of each word in the image is an warning of frequency of occurrence of the word in the entire text. We turn the corpus into a structured data by processing the texts by removing the numbers, stop words, punctuation, Unicode characters, making all the words uniform to lowercase and finally, unnecessary white spaces are stripped. Now, the frequency of every word is calculated in the resultant data, sorted with respect to its frequency of occurrence, converted to a data frame and a cluster of words is formed.

3.2.6 Visualization of Top Most 10 Frequencies of Occurrences of the Words in the Analysis

From the very same data processed for creating the word cloud, we visualize the top most 10 frequency of occurrences of the words by extracting the top 10 words with its frequency in the created data frame with "rCharts" library.

3.2.7 Visualization of Location of the Tweets

Location of the tweets for a particular movie is tracked down by prompting the user to enter the latitude, longitude pair of location along with the radius (a coverage actually) until which the location of tweets is searched for. It is visualized with the help of "leaflet" package. It is designed in a way that if the user places the pointer over the marker on the location, it will display the name of the twitter user who tweeted the tweet.

3.2.8 Integrating All the Components in Shiny RStudio Dashboard

Shiny RStudio is a web application framework for R. It helps take the R script to the next whole new level. Instead of running the script that are scattered here and there, Shiny provides a space for creating an application in R under one roof with an elegant UI. It consists of two R scripts 'ui.R' and 'server.R'. 'ui.R' consists of code written to interact with user and all those values are passed to the server side in 'server.R' like all other programming languages. 'server.R' consists of the script for computation, processing and visualization of data. If the developer wishes, he can take the application to be hosted on Shiny server online, so that the users worldwide can utilize the application and make use of it. Fig. 3 represents the process of integrating all components in RStudio Dashboard.

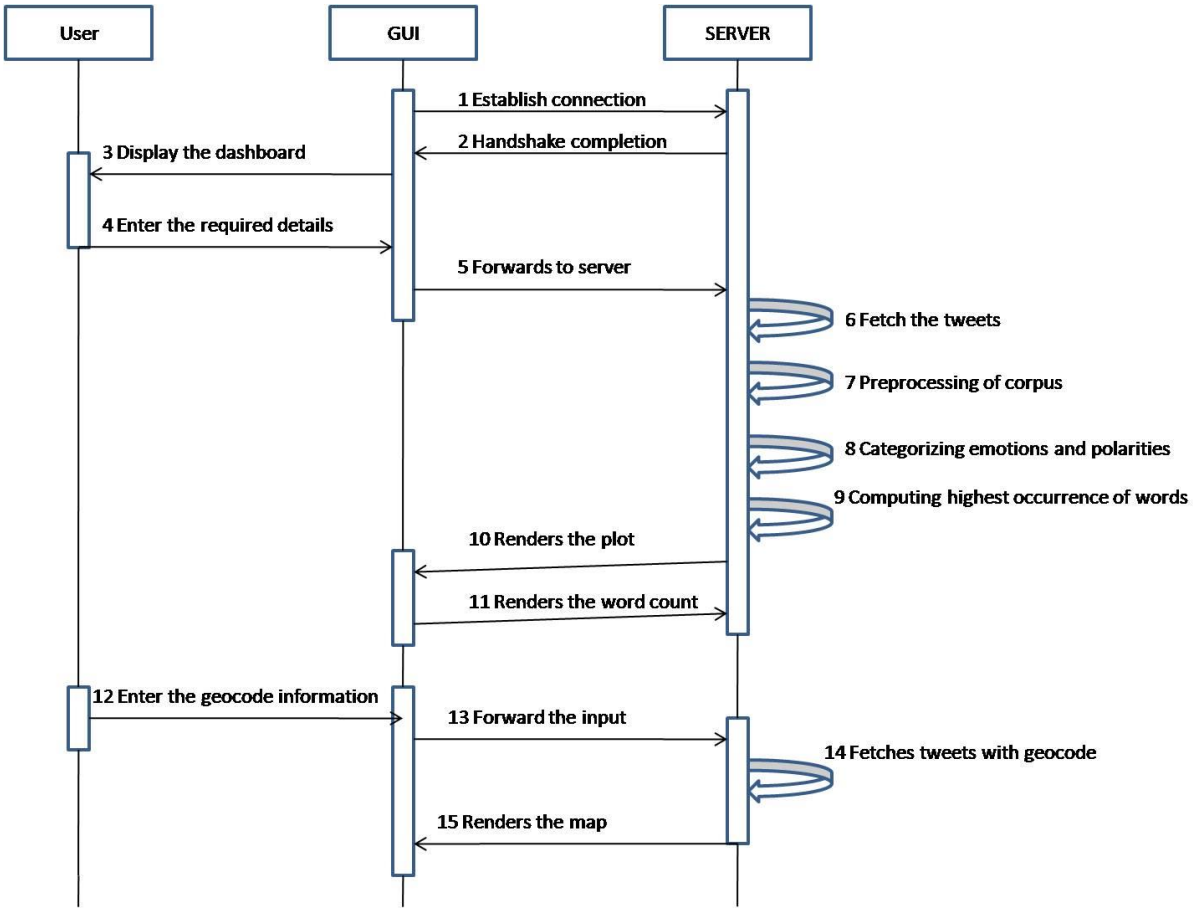

Fig 3. UML Sequence Diagram for Integrating into Dashboard

4. Performance Evaluation

The proposed L2B model is validated based-on using four metrics such as accuracy, sensitivity (recall), positive predictive value (precision) and F1-measure. A set of these four metrics have been applied for every set of training and testing. The proposed model uses Naive Bayes Classifier supervised Machine Learning text classification algorithm, Lexicon approach and has achieved nearly 86% of the accuracy. A total of 10944 rows

of text were made to undergo training and testing out of which 7660 rows of text were used for training and 3284 rows of text were used for testing. We attained this accuracy after days of training started from few sets of data, gradually increased and ultimately ended up with this result. Quality (structure) of data is also one factor that determines the accuracy of the model.

Table 1. Performance evaluation of ETPP

Total no. of rows of text	No. of rows of text for training	No. of rows of text for testing	Accuracy (%)
150	105	45	89.2
600	420	180	89.0
1500	1050	450	89.4
4000	2800	1200	88.2
6000	4200	1800	85.2
7500	5250	2250	87.8
9000	6300	2700	86.1
10944	5660	3284	86.4

A snippet of the performance evaluation is shown below:

```
# measuring Naive Bayes classifier's performance # training the model
>classifier = naiveBayes(mat[1:7660,], as.factor(smpl[1:7660,2]))
# testing the model
> predicted = predict(classifier, mat[7661:10944,])
>recall_accuracy(smpl[7661:10944,2], predicted) [1] 0.8636179
>confusion.mat=table(predicted,(smpl[7661:10944,2],dnn= list('predicted','actual'))
> confusion.mat
predicted actual       negative positive negative 1416      230
positive     229        1409
```

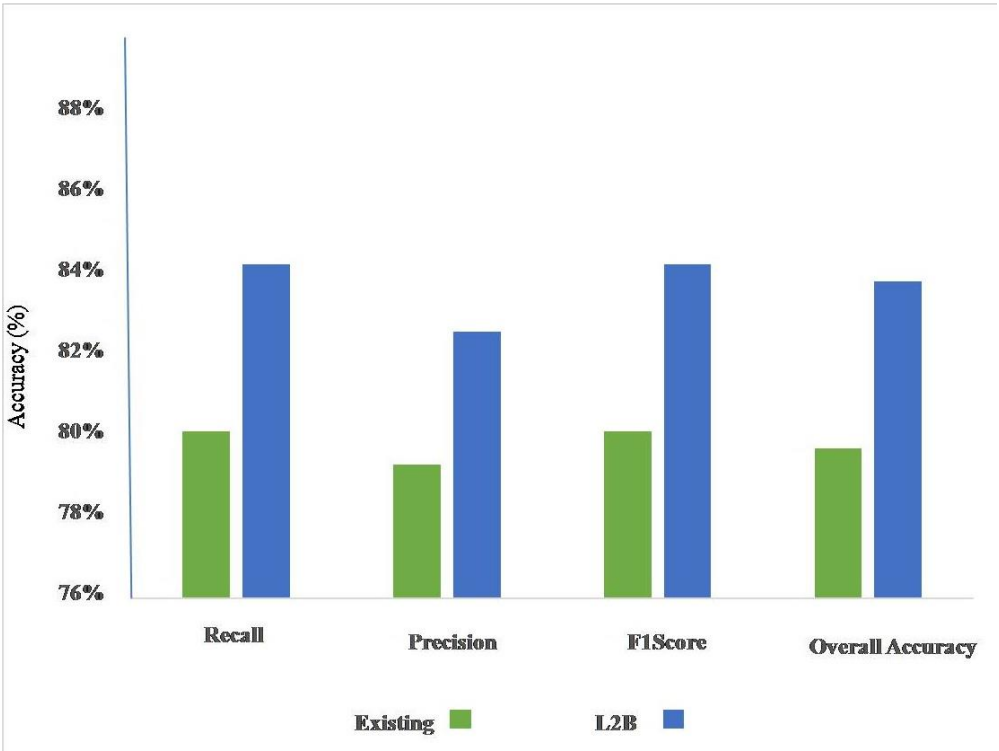

Fig.3. Comparison of L2B with existing approach for Accuracy

5. Conclusion and Future work

People would find this proposed work, an application helpful as it crunches the huge opinions of tweeple about the movie in numbers and emotions that reflect the state of mind of them. It has been designed to predict

movie popularity with "Naive Bayes Model" and "Lexicon approach" that gives a cumulative accuracy of 84% greater than what the existing models. Proposed work contains only the retrieval of tweets only in English language because if multiple languages were preferred, training would have been more difficult. Hence, this paper could be extended to retrieve tweets in languages other than English (enabling the proposed work to be localized stronger) and train the model. Accuracy of the Naive Bayes Model can be achieved higher by incorporating training by using large the dataset. Making the RScript to work on a regular basis using "Task ScheduleR" in R which enables dynamic retrieval of data without human intervention and it ultimately plots the data that will dynamically keep on changing until the user searches for another topic or quits the application.

References

1. Aryan G, Vinya D, Hamza K & Manan S, "Comprehensive review of text mining applications in finance, Journal of financial innovation, pp. 1-25, 2020.
2. Felix M F W, Soumya S & Mung C, "Why watching movie tweets won't Tell the whole story?", Princeton University, New Jersey, 2012.
3. Dubey P K & Agrawal S, "A critical analysis of twitter data for movie reviews through random forest approach", International conference on information and communication technology for intelligent systems, vol. 2, pp. 454-460, 2017.
4. Domingos P & Pazzani M, "On the optimality of the simple Bayesian classifier under zero-one loss", Machine Learning, vo. 29, pp. 103–130, 1997.
5. Tholkapiyan, A.Mohan, Vijayan.D.S, A survey of recent studies on chlorophyll variation in Indian coastal waters, IOP Conf. Series: Materials Science and Engineering 993 (2020) 012041, 1-6.
6. Dipak D, Gaikar & Bijith, "Using twitter data to predict the performance of bollywood movies", Emerald Groups Publication, vol. 115, no. 9, pp. 1604-1621, 2015.
7. Jain V, "Prediction of movie success using sentiment analysis of tweet", International journal of soft computing and software engineering, vol. 3, no. 3, pp. 308-313, 2013.
8. Liangxiao J, Harry Z & ZhihuaCai, "A novel bayes model, hidden naive bayes", IEEE Transaction on knowledge and data engineering, Vol 21, No. 10, pp. 1361 – 1371, 2009.
9. Mohammed M F, Tarek F G & Abdulfatah S M, "Efficient twitter sentiment analysis system with feature selection and classifier ensemble", International conference on advanced machine learning technologies and applications, pp. 516-527, 2018.
10. Gopalakrishnan, R., Mohan, A., Sankar, L. P., & Vijayan, D. S. (2020). Characterisation On Toughness Property Of Self-Compacting Fibre Reinforced Concrete. In Journal of Environmental Protection and Ecology (Vol. 21, Issue 6, pp. 2153–2163).
11. 229, Fall 20Nan L & Desheng D W, "Using text mining and sentiment analysis for online forums hotspot detection and forecast", Elsevier Publication, 2010.
12. Nirmala Devi K & Murali Baskaran V, "Sentiment analysis For online forums hotspot detection", ICTACT Journal on soft computing, vol. 02, no. 02, 2012.
13. Gopinath, S., Rajaram, A., & Suresh Kumar, N. (2012). Improving minimum energy consumption in ad hoc networks under different scenarios. *International Journal of Advanced And Innovative Research (IJAIR), 1*(4), 40-46.
14. Ravi K & Ravi V, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications", Knowledge based systems, vol. 89, pp. 14-46, 2015.
15. Sitaram A & Bernardo A. Huberman, "Predicting the future with social media", International conference on web intelligence and intelligent agent technology, vol. 1, pp. 492-499, 2010.
16. Sameer T & Tushar P, "Prediction of box office success of movies using hype analysis of twitter data", International journal of inventive engineering and sciences, vol. 3, no. 1, 2014.
17. Vasu J, "Prediction of movie success using sentiment analysis of tweets", The International journal of soft computing and software engineering publication, vol. 3, no. 3, 2013.