International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 04 | Apr 2025 www.irjet.net p-ISSN: 2395-0072

Design an Approach for prediction of Popular Movies Based on Previous Reviews by Using Bagging & Boosting Techniques

Bharti Singh [1], Prof. Manish Rohila[2]

MTech Scholar, Department of Computer Science & Engineering, Technocrats Institute of Technology (Excellence), Bhopal (M.P), India, Email: bhartisingh15193@gmail.com
Assistant Professor, Department of CSE, Technocrats Institute of Technology (Excellence), Bhopal (M.P), India[2]

ABSTRACT

Here, Authors explained their views with the help of experimental events & setup. In this work the selected data set is movies reviews given or collected by concern people. This data has taken from the nltk.corpus package defines a collection of corpus reader classes, which can be used to access the contents of a diverse set of corpora. NLTK's corpus reader classes are used to access the contents of a diverse set of corpora. Each corpus reader class is specialized to handle a specific corpus format. The dataset selected as previous which we took as our Base Reference for our research. the dataset selected having two categories negative & positive each category contains 1000 files. In our file we have movies review by different reviewer. By the given dataset we need to extract sentiment from the given text documents. we extend some more in comparison to our base papers. we apply Bagging & Boosting in sentiment analysis tool, which is basically a way to figure out if a piece of text is expressing positive, negative, or neutral emotions. VADER concept based upon Bags of Words approach. Authors used a concept Bagging & Boosting which comes under Ensemble Techniques. This concept improves the performance by 1 % in voted algorithms.

Keywords: Sentiment Analysis, Stop Word, Tokens, Features, Training & Testing Data, Model or classifier, VADER, Ensemble Learning, Bagging, Boosting

I INTRODUCTION

In today's digital era, social media platforms play Vital roles for transforming the individual's life, they share information in the public domain and try to interact with different people and organizations. The process of collecting, analyzing, and finding the insight from these given data, by this insight we are able to find any one's behaviour and trends. Now days social media plays very critical role of finding any analytics with the help of Data science. For finding these we need number of Tools & techniques. because by using these tools we can find the trends in effective manner [1].

In this computer world every company and form want that reviews of their products goes in market with the help of print media and social media because the positive review can pull many new consumers to them without much expenditure in their marketing plans. If you see major FMCG companies depends upon their products feedback by the customers who purchase their product [2].

we all knows that this decade is digital decade all different media are not so popular in comparison to social media. Even though many big online companies depend upon social media advertisement. as example if we go through the companies like Zomato then we find that Zomato is now very big company in comparison to many FMCG Companies. If we go to in depth then we find that how uber gives the many chances to their drivers who have very good reviews given by past customers. All the above example comes under sentiment or opinion of customers after availing the services by different providers from many different Domains [3].

1.1 Machine Learning Raw Sketch

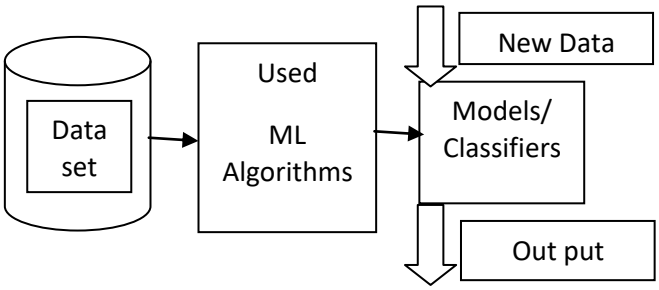

Figure 1: Machine Learning Raw Sketch [3]

In figure:1 Authors explain the very first concept of machine learning here the very first part is dataset where concern data has been stored. Now then we pass the part of data to next part that is known as training, in this part we pass the data to a given ML Algorithm that create a specific Model or classifiers that will used further for prediction a new data sample directly given to the model.

Machine learning is broadly categories into following based upon Dataset behaviour:

Labelled Data: If Label available in Dataset: Any Machine learning applied on labelled Data that comes under Supervised Machine Learning [4].

If o/p is continuous → Regression Project

Regression: It predicts continuous values

Ex: Salary, Price, Sales

Else o/p is discrete → Classification Project

Classification: Predicting the item class / category of a case

i) Binary Classification Project

ii) Multiclass Classification Project

Unlabelled Data: if label not available in Dataset: Any ML techniques applied on Unlabelled Data that comes under Unsupervised Machine Learning.

Clustering: Finding the similar item at a single place.

Association Rules: Associating frequent co-occurring items /events

Ex: Laptop + Bags, Bread + Butter + Jam

Recommendation Engine: It gives recommend based upon your previous interest.

1.2 Types of Datasets

Corpus Reader Objects ("Corpus Reader Objects") describes the corpus reader instances that can be used to read the corpora in the NLTK data package.
Most corpora consist of a set of files, each containing a document (or other pieces of text). A list of identifiers for these files is accessed via the fields () method of the corpus reader. Each corpus reader provides a variety of methods to read data from the corpus, depending on the format of the corpus [5,6].

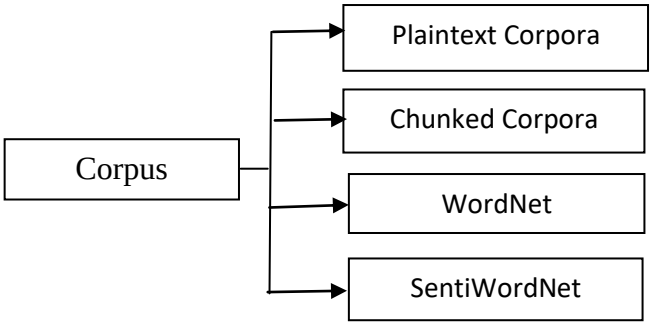

Figure 2: NLTK Corpus Data Types [5,6]

Plaintext Corpora: It contains plain text information in form of files that can be process with the help of nltk methods.
Chunked Corpora: It also provides chunk structures, which are encoded as flat trees.

WordNet: WordNet is kind of lexical database storage or you can say dictionary for the English language, specifically designed for natural language processing.
SentiWordNet: It provides the functionality like the measure of positivity, negativity or neutrality which is needed for Sentiment Analysis.

1.3 Required Framework for Sentiment Analysis

For working with sentiment analysis, we have to download nltk module. In this module we have to download all the required folder. We have to follow following steps to install all required modules [7].

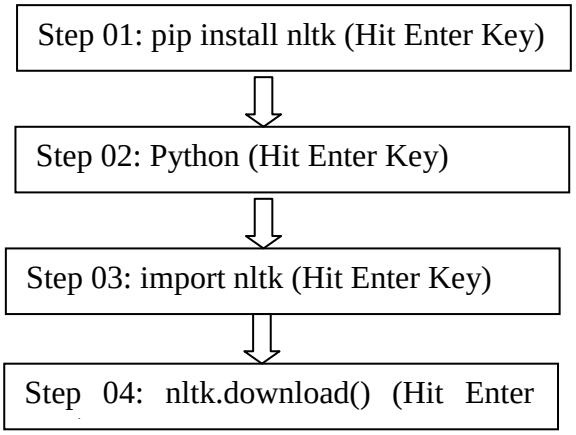

Figure 3: Process to Download sentiment Module [7]

In Figure 3: Authors explain how the nltk tool has been install at your personal system for further analysis. The above commands will play at command prompt.

1.4 Ensemble Learning & Voted Algorithm

In Ensemble learning we look at multiple classifiers and combining the output of multiple classifiers in order to get better prediction or classification or classification accuracy [8-11].

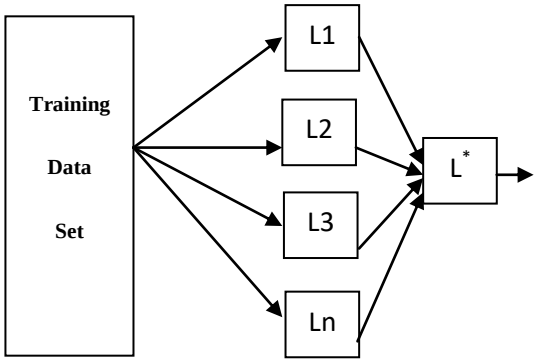

Figure 4: Ensemble Mechanism

Explanation: In above figure Training Data set may be different Data set or may be split into different Data set. We have different Base Learners or model like L1, L2, L3, …

…. Ln. Here every learner can use same algorithms with different parameters and different algorithms like naïve bayes, SVM, LR and many more. Finally, we have to combine using voted mechanism where we can use hard voting as well as soft voting.

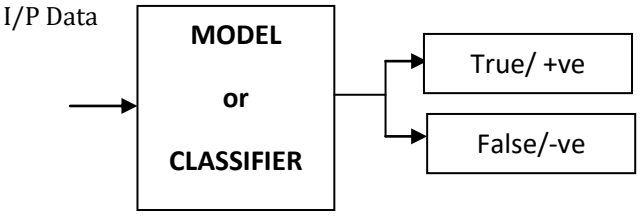

Figure 5: 2 - class classifier

Explanation: Ensemble use multiple trained models means combining weak learners and give you strong one. Assume n independent learners out of them:

N_1 say class1
N_2 say class2
N_3 say class2
.
.
.
N_N say class1

By using voting algorithm, we will claim which class has majority opinion.

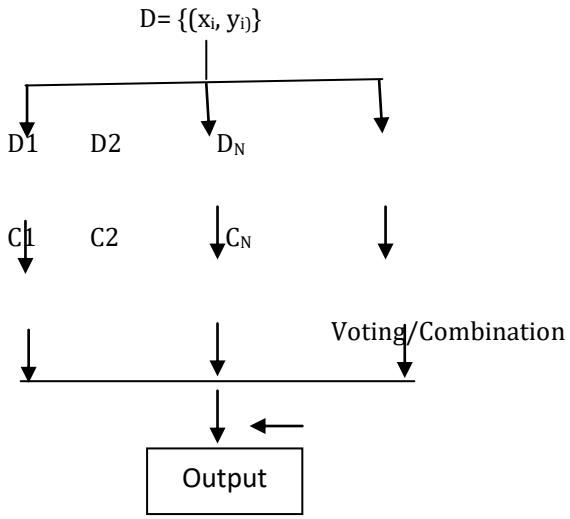

Figure 6: Voting Mechanism

1.5 Bagging & Boosting

Bagging: Bagging =" Bootstrap Aggregation", use bootstrapping to generate L training sets and train one base-learners with each.

Boosting: In order to use an ensemble of learners and able to combine the learners and getter better accuracy, every learner make independent error.

Note: If you use a learner on a small training set the learner can overfit, it can have high variance and it will not generalize [12].

II RELATED WORK

In this paper the authors used A collection of movie reviews is what provides us with a deeper qualitative insight on various aspects of the movie, whereas providing a movie with a numerical rating in the form of stars tells us about the success or failure of the movie quantitatively. We can learn about the movie's strengths and weaknesses from a textual review, and a more in-depth analysis of a movie review can tell us if the movie overall meets the reviewer's expectations. In this project, we want to use Sentiment Analysis on a set of movie reviews written by reviewers to figure out how they felt about the movie overall, such as whether they liked it or hated it. We want to use the relationships between the words in the review to predict the review's overall polarity [13].

In this paper the author, Movie reviews are vital in telling the viewer whether a movie is worth watching or not. They can be classified into textual and non-textual movie reviews. While non-textual movie reviews (stars) give the user information as to how the movie fairs, textual movie reviews give the user a more detailed picture on the positive and negative aspects of the movie. This paper aims to conduct sentiment analysis of reviews of movies by using the Naive-Bayes algorithm and compare the results to that of a Rule-Based Approach using the AFINN-111 sentiment dictionary. [14]

In this paper the authors proposed a system in which it has been assured that this is the one Act has been initiated by the Government of India in September 2020, often known as Farm Bills or Indian Agricultural Act. This act has been affected farmers in many ways and led to opposition to the bills. As a result, there is a wide area for doing sentiment on the data taken from this domain, so we are making sentiment analysis on it. On comparing different algorithms like Logistic Regression, VADER, and BERT we could see that BERT is having more accuracy as compared to the other algorithms. But we could see that VADER is a good algorithm as they are having special qualities as compared to that of the other algorithm. So, we thought to Improve VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis on Farm Bill Act. Along with this, we are doing Information Extraction on verb and analyzing sentiments on extracted phrases that are related to the verb, to get the accuracy of sentences with verb and without verb. Thus, we get the Dependency of the verb in the sentence. [15].

III PROBLEM IDENTIFICATION

A lot of research work has been done in the field. Authors have learned several things from this study (work).

Authors find that in sentiment analysis has been done by many authors but percentage of positive and Negative is missing by many authors in their research. Then finally Authors Decide that they will try to implement **VADER** (**Valence Aware Dictionary and Sentiment Reasoner**) is a rule-based sentiment analysis tool. That will give us more accuracy than previous one. Authors also implement the previous methods used by researchers. They change some mechanism like Bagging & Boosting the Existing Ensembled Techniques with Voted Algorithms.

IV ALGORITHMS

Step 01: Select our Data set where we will select Data

In this section we fetch our Data Set Movies Review from Corpora Community. In this Data we have 2000 files containing 1000 negative review & 1000 positive review.

Step 02: Perform data cleaning over that data and extract keywords from Data Set

- **(i)** Tokenization
- **(ii)** Stop Word Finding
- **(iii)** Root word Finding
- **(iv)** POS (Parts of Speech)
- **(v)** Frequent Word Finding

Step 03: Calculate Frequency count

Step 04: Apply **VADER** (Valence Aware Dictionary and Sentiment Reasoner)

Step 05: Apply ML Algorithms

For Naïve Bayes

- **(i)** Import Libraries like pandas, NumPy, sklearn
- **(ii)** Assigning predictors & Target Variable
- **(iii)** Apply Model like NB
- **(iv)** Fit the model
- **(v)** Predict the Output
- **(vi)** Compare Result Predicted & Actual

Step 06: Repeat step 05 multiple times for every classifier (make sure the parameters used for every algorithm may change)

Step 07: Calculate Different performance parameters like Accuracy, Root Mean Square, F1-Score and many more.

Step 08: Repeat the Step 07 & compare to find which gives best Result

Step 09: Stop

V ARCHITECTURE OF SYSTEM

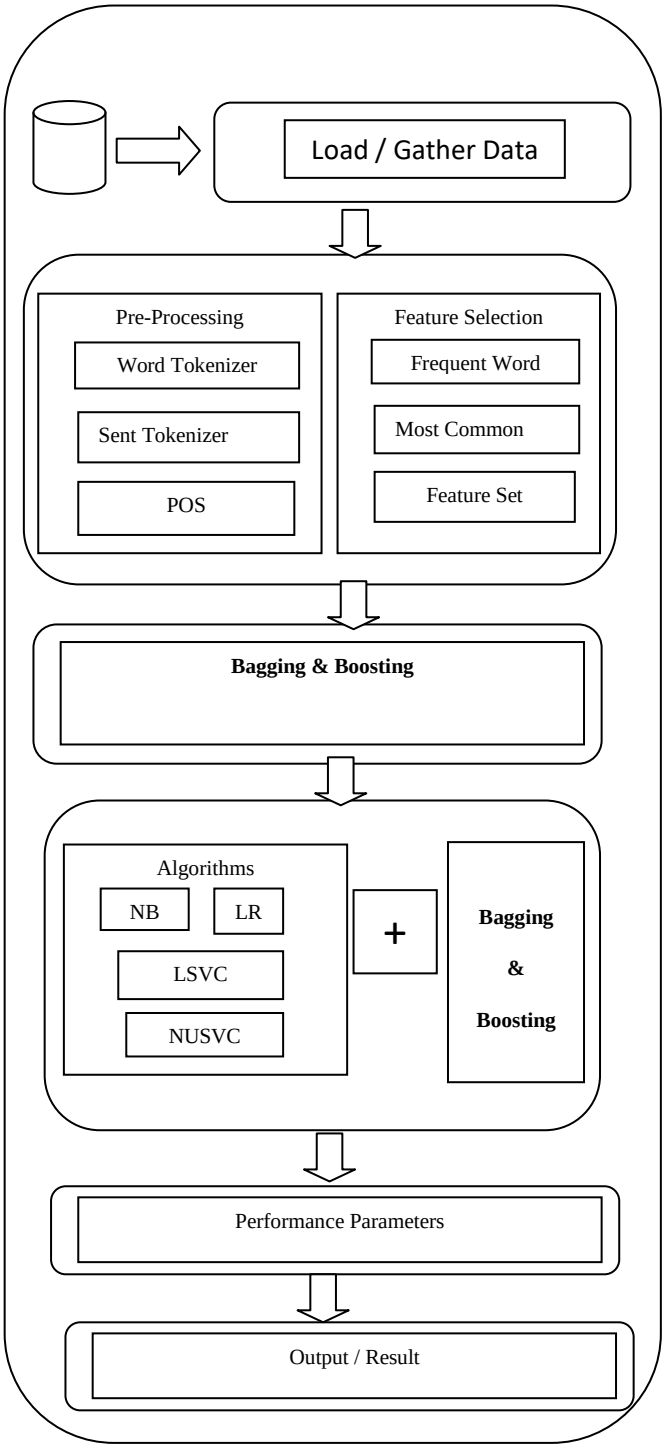

Figure 7: Flow Diagram

In Figure:7, Authors Explained that Data is taken from Data Source then the part of data send to next phase i.e. Preprocessing and Feature selection mechanism. In Preprocessing taken sentence is broken into word and token that can be process easily in nltk. In Feature selection they picked most frequent word that help for

finding for feature set. Here they implemented new concept here i.e. VADER, which gives us positive, negative & neutral between a Given band i.e. (-1 to +1). After these processing they implement previous Algo as well as some change concept i.e. Bag of words that comes from VADER. This concept will implement in Ensemble techniques with Bagging & Boosting. Finally, they find all required performance parameters and put it into table for comparison and try to show which one is better fir business use during End-to-End Implementation.

VI IMPLEMENTATION AND RESULT

Explanation: After analysing the above table, we can say that with variations in features in our data set our results is changing. Results also depends upon Training and Testing split values. We analyse the results only on Naïve Bayes we can do this on different Algorithm also. it may show a sharp and definite pattern of our analysis.

Comparison of Different Algorithms					
S.No	**Algorithm**	**Features**	**Training**	**Testing**	**Accuracy**
1	**Bagging & Boosting (Proposed Algorithm)**	**3000**	**1900**	**1100**	**92**
2	**MNBNB**	**3000**	**1900**	**1100**	**86**
3	**BernauliNB**	**3000**	**1900**	**1100**	**84**
4	**Logistic Regression**	**3000**	**1900**	**1100**	**89**
5	**Linear SVC**	**3000**	**1900**	**1100**	**83**
6	**NUSVC**	**3000**	**1900**	**1100**	**85**

Table 01: Comparison of Different Algorithm

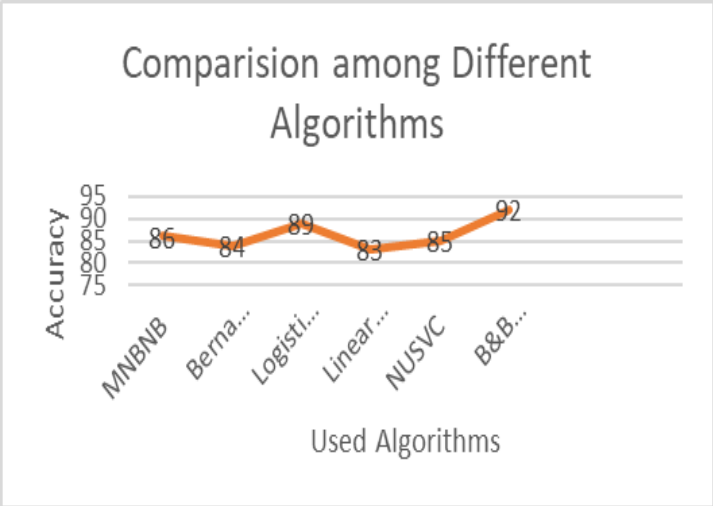

Figure 8: Line Graph Between Algorithms

Explanation: In the above image we can say that B&B I.E Bagging & Boosting algorithm is giving better performance.

S.No	Algorithm	Previous Work (Accuracy %)	Proposed Work (Accuracy %)
1	**MNBNB**	**86**	**-**
2	**Bernoulli**	**84**	**-**
3	**Logistic Regression**	**89**	**-**
4	**LSVC**	**83**	**-**
5	**NUSVC**	**85**	**-**
6	**Bagging & boosting**	**91**	**92**

Table 02: Analysis of previous work & proposed work

Explanation: In the above Table we explained the percentage Accuracy between previous work and proposed work. Naïve bayes & Random Forest gives better result but SVM gives low Accuracy percentage. In this work authors used Bagging & Boosting Mechanism that gives us better result. Authors attain 92 (percent) in Bagging & Boosting Algorithms.

Note: Bagging & Boosting is extension of voted Algorithms which comes from Ensemble techniques.

VII CONCLUSION

Here Authors implement their model even though they feel that in sentiment analysis many authors did much work but still there is much scope in performance improvement. And during their model implementation they found that features splitting is very key points in their dataset movies

review. They added a new concept i.e. **Bagging & Boosting** that improves during analysis because it gives the how much positive and Negative sentiment they receive. During Model creation they used Bagging and boosting mechanism that improves the quality by +1 %.

VIII FUTURE SCOPE

During study Authors find that sentiment analysis has major three components Positive, Negative and Neutral. But the given result improves with the help of VADER. In near future BERT can also gives better result in this case. If we try another different algorithm that may also increase in performance point of view. In near future complex sign also including during sentence posting that can also to huge task to complete for Data scientist.

REFERENCES

[1] Rubeena Parveen, Neelesh Shrivastava and Pradeep Tripathi, “Sentiment Classification of Movie Reviews by Supervised Machine Learning Approaches Using Ensemble Learning & Voted Algorithm”, 2nd International Conference on Data, Engineering and Applications (IDEA), IEEE

[2] Navaneetan M, Tharagesh G, A Sai Sabitha " Unveiling Sentiments: Analyzing Learner’s experience using VADER and RoBERTa models ", 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems, IEEE.

[3] Phumelele P. Kubheka, Pius A. Owolawi, " Harnessing Advanced Classifiers for Sentiment Analysis on Augmented Tweets”, IEEE.

[4] Nadimpalli Madana Kailash Varma, Sri Harsh Mattaparty, “Sentiment Analysis: A Machine Learning perspective", 2024 First International Conference on Electronics, Communication and Signal Processing (ICECSP) | 979-8-3503-6459-0/24 ©2024 IEEE | DOI: 10.1109/ICECSP61809.2024.10698402

[5] Tirath Prasad Sahu, Sanjeev Ahuja,"Sentiment Analysis of Movie Reviews: A study on Feature Selection & Classification Algorithms",978-1-4673-6621-2/16 2016 IEEE

[6] upma kumari, Dr. Arvind k. Sharma, Dinesh soni,"sentiment Analysis of smart phone product Review using SVM Classification Techniques" International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017).

[7] Rasika Wankhede,Prof. A.N.Thakare,"Design Approach for Accuracy in Movies Reviews Using Sentiment Analysis",International Conference on Electronics, Communication and Aerospace Technology ICECA 2017

[8]Kamil Topal,Gultekin ozsoyoglu,"Movie Review Analysis: Emotion Analysis of IMDb Movie Reviews",2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

[9]Akshay Amolik, Niketan Jivane, Mahavir Bhandari, Dr.M.Venkatesan ,"Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques",e-ISSN : 0975-4024 Akshay Amolik et al. / International Journal of Engineering and Technology (IJET)

[10] Rahul Chauhan, Aman Gusain, “Fine Grained Sentiment Analysis using Machine learning and Deep learning”, (ICSEIET) , 2023