IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

NEWS CLASSIFICATION AND RECOMMENDATION USING NAIVE BAYES CLASSIFIER

[1]Himanshu B. Chaudhari, [2]Mayur G. Kotkar, , [3]Aditya S. Pardeshi, [4]Pratik V. Ukarde, [5] Prof. S. N. Chaudhari

[1,2,3,4] Students, Department of Information Technology Engineering

[5] Professor, Department of Information Technology Engineering

K. K. Wagh Institute of Engineering Education and Research, Nashik, India Affiliated under Savitribai Phule Pune University

Abstract: News publishers have decreased spreading news through conventional newspapers and have migrated to the use of digital media. Therefore, there exists a large amount of information being stored in the electronic format which needs to be classified into different categories because there may be present some sensitive data which is not suitable for specific age group. In this project, machine learning algorithms are used for classifying news by using a dataset. By evaluating the accuracy of Linear regression, Naive Bayes algorithm, Logistic regression and Decision tree algorithm and performing comparative analysis of all mentioned algorithms we are going to select the algorithm which provides the maximum accuracy. Also suggesting news articles to the online news reader based on the similarity of a news article with the news they are reading or proposing news articles based on the interest of news reader subjected from their previous readings and the feedback of the reader. After implementing and comparing the accuracy of all Machine Learning models, Naive Bayes Model gave the highest accuracy and lower error percentage. So, Naive Bayes Model is used for training the model and to get required result.

Index Terms **- Naive Bayes algorithm, Comparative analysis, Decision tree algorithm.**

I. INTRODUCTION

The advancement in technologies like high-speed internet access on handheld devices and customized software applications for news is beneficial for news readers as they provide very easy access to the information published on multiple sources. It is the human inability to browse through such a vast information space for related news stories. The number of news releases has grown rapidly and for one individual, it is difficult to browse through all online news resources for relevant news articles. Search engines help users up-to some extent in searching through the vast information collections available online and the recommendation systems have emerged to address different challenges and provide users with the information which matches their needs either by their preferences or by content similarity. Each online news publisher tries to handle its news and use some mechanisms to recommend similar news to their readers[1].

An application that helps to predict items based on the user responses to other options of items set or a meaningful recommendation to a collection of users for product or items that may interest them is a recommendation system. The recommendation systems help users to find information and make decisions where they lack the required knowledge to judge a particular product. Also, the information dataset available can be huge and recommendation systems help in filtering this data according to users needs. For example, suggesting news on online newspaper website using recommender systems. However, sometimes the news suggested by recommendation system is not good to read for specific age group. Here, this machine learning model will classify the news based on specific age group and recommend the news to the user.

The objectives section of the paper aims to propose a method for classification and recommendation of news in machine learning that can successfully show the news to user based on its interest without compromising the model's performance quality. It explains the importance of recommendation of news which fulfills the user interest.

The literature section of the paper covers various topics related to news classification, various machine learning algorithms like naive bayes classifier, and recommendation-based learning. It includes discussions on Recommendation of news and its benefits, its drawbacks, challenges to implement the recommendation system, naive bayes algorithm. Also covers the solutions to increase the accuracy of model to generate the accurate results.

The methodology section provides the information about the problem statement and the importance of naive bayes classifier. Naive Bayes classifiers are a collection of classification algorithms based on Bayes Theorem. Naive Bayes classifier is explained with its block diagram and its model. Also their is a high level presentation and design of project using various UML Diagram like class diagram, use-case diagram and activity diagram.

The results section presents the findings of the study, using tables, figures, and charts to illustrate the data analysis. It provides a clear and concise summary of the experimentation. The conclusion summarizes the main findings of the study and restates the research question.

II. OBJECTIVES

The primary objective of this experiment is to evaluate the effectiveness of various machine learning algorithms to classify the news based on user interest and its age while ensuring that the high machine learning model's accuracy is achieved. To do this, the available machine learning model should be examined for news classification by comparative analysis and a method should be suggested that can provide high accuracy of classification without compromising the model's performance quality.

Recommendation is crucial in information retrieval because it can shorten the time of finding the news. The accuracy of the machine learning model may be impacted, if crucial data is lost or duplicate data is present in dataset. Eventually it affects the quality of news which is classified and recommended by model.

To solve this problem, there is a need to concentrate on finding of a such machine learning algorithm that can provide the high accuracy of model. This method will be created by researching current methods and determining where advancements can be made.

III. LITERATURE SURVEY

This section covers a broad review of dataset reduction techniques, use of clustering algorithms for dataset reduction in particular, the formation of homogenous clusters.

According to Chong Feng, Muzammil Khan, Arif Ur Rahman And Arshad Ahmad, News publishers have decreased spreading news through conventional newspapers and have mi- grated to the use of digital means like websites and purpose-built mobile applications. It is observed that news recommendation systems can automatically process lengthy articles and identify similar articles for readers considering pre defined criteria. The objectives of the current work are to identify and classify the challenges in news recommendation domain.

However, The recommendation of news articles is a hard task because of a highly dynamic environment, which leads several challenges, e.g. frequent changes in the set of news articles, set of users, rapid changes in users preferences, etc. Therefore, recommendation algorithms must be able to process continuous incoming news streams in real-time[1].

According to Mr. M. Sundarababu, Ch. ChandraMohan, Mahendra Suthar, Ch. Deva Harsha, Lubna Juveria and B. Blessy, There exists a large amount of information being stored in the electronic format. With such data, it has become a necessity of such means that could interpret and analyze such data and extract such facts that could help in decision making. Data mining which is used for extracting hidden information from huge databases is a very time consuming process.

However, The major drawback would be its provided accuracy. The second drawback is Zero Frequency Problem that exists in the naive Bayes algorithm. Therefore, to overcome these drawbacks we are using the Multinomial Naive Bayes Algorithm[2].

According to [Das et al., 2007]. Here, both model and memory based algorithms have been used. Memory based algorithms use weighted average of past ratings from other users where weight is proportional to the 'similarity' between users. Pearson correlation coefficient and cosine similarity are the typical measures used for 'similarity'. Model based ratings model the user preferences based on past information. From model based, they used clustering techniques PLSI and MinHash, and from memory based, they used item covisitation. All the three algorithms assign a score to a story. Finally, two combinations of three techniques (PLSI, MinHash and co-visitation) were used for evaluation. In one combination, co-visitation was given a higher weight than the other techniques (2.0 instead of 1.0). This method was named CVBiased and in another combination PLSI and MinHash were given higher weight (2.0 instead of 1.0), which is known as CSBiased. The baseline usedwas recommendation based on recent popularity called Popular. It was observed, on average both CVBiased and CSBiased performed 38 percent better than the baseline Popular on live traffic[3].

According to [Li et al., 2010]. Here, news recommendation problem was modelled as a contextual bandit problem. Authors devised a new algorithm named as LinUCB to solve the problem. LinUCB is a general contextual bandit algorithm and can be applied to domains other than news recommendation. Using LinUCB, number of clicks increased by 12.5 percent as compared to a standard context-free bandit algorithm[4].

IV. PROBLEM STATEMENT

People tend to have preferred sources for printed news, like a newspaper or a magazine. However, in online world, this reality changes drastically, a user is flooded with information from different sources in which it is not uncommon for a user to change between news portals or read news from portals that merge news articles from different sources (Google News). Having such a huge amount of information, it becomes difficult to select the news article that a user will like. Consequently, users stop the consumption of news or lowers it down. To overcome the problem of information overloading (difficulty in making a decision caused by the presence of too much information), recommender systems can be used. As on the Internet, there exists a large amount of information being stored in the electronic format which needs to be classified into different categories because there may be present some sensitive data which is not suitable for everyone. So, there is need of classification of news in accordance with specific age group.

V. METHODOLOGY

Naive Bayes classifiers are a collection of classification algorithms based on Bayes Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Naive Bayes like logistic regression predicts if an input belongs to a particular category or not. It's output is also probability but instead of logistic function, it uses Bayes rule. Following figure describes Bayes rule appropriately :

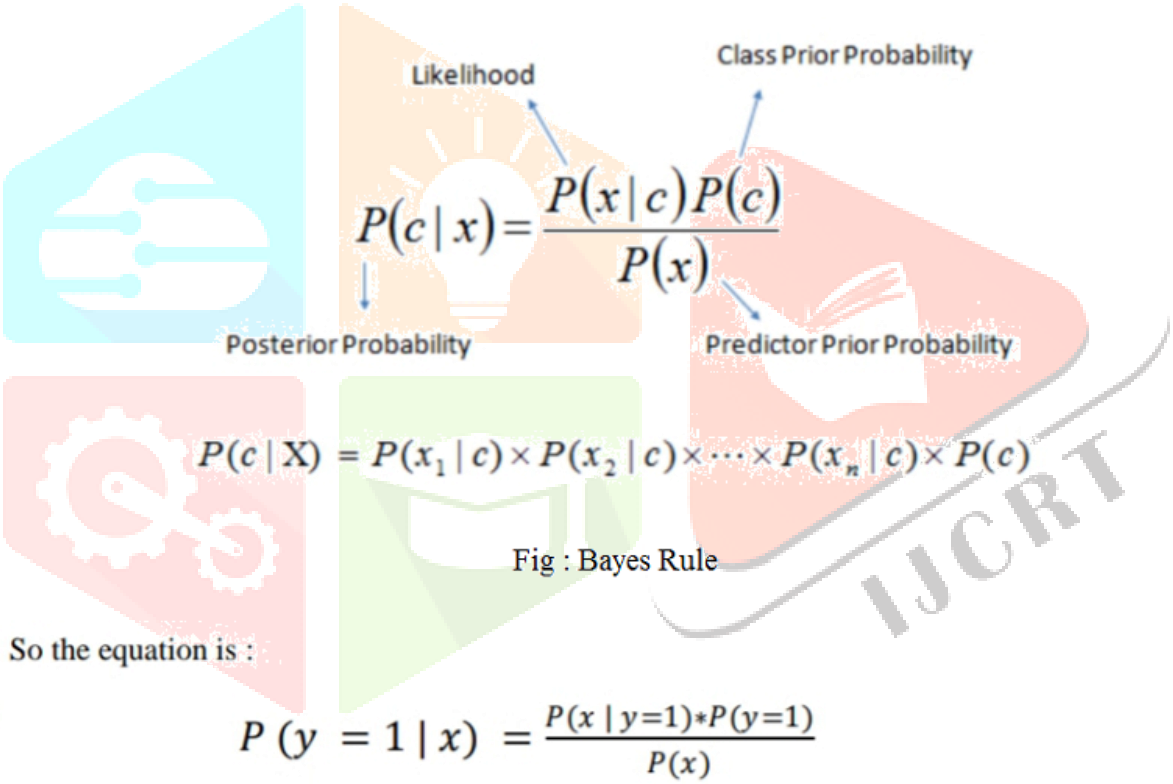

Fig : Bayes Rule

So the equation is :

$$P(y = 1 \mid x) = \frac{P(x \mid y=1) * P(y=1)}{P(x)}$$

where y is output and x is input. P(x | y) and P(y) are calculated from the given data and used for predicting outputs of new inputs. For different values of y, P(y | x) is calculated and x is assigned to the category (or y is assigned the value) for which P(y | x) is maximum. An assumption is made that variables x1,x2.... are independent of each other and therefore, P(x1,x2.... | y) can be written as P(x1 | y) * P(x2 | y)....., which is not true is most cases and therefore this algorithm is called Naive Bayes[5].

The trick here is Machine Learning which requires us to make classifications based on past observations (the learning part). We give the machine a set of data having texts with labels tagged to it and then we let the model to learn on all these data which will later give us some useful insight on the categories of text input we feed[6].

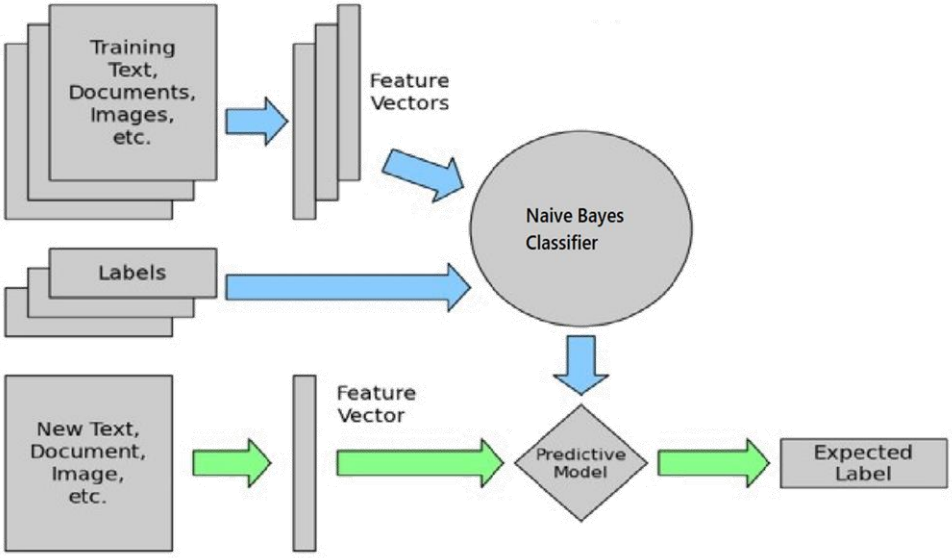

Fig : Block Diagram

It works on the famous Bayes theorem which helps us to find the conditional probabilities of occurrence of two events based on the probabilities of occurrence of each individual event.

VI. RESULTS AND DISCUSSION

The experimentation is done on a News Headlines India dataset.

DATASET INFORMATION :

This news dataset is a persistent historical archive of notable events in the Indian subcontinent from start-2001 to q1-2022, recorded in real-time by the journalists of India. It contains approximately 3.6 million events published by Times of India.

Content :
Time Range : Start Date : 2001-01-01 ; End Date : 2022-03- 31
CSV Rows : 3,650,970

Columns :
1. publish date : Date of the article being published online in yyyyMMdd format.
2. headline category : Category of the headline, ascii, dot delimited, lowercase values.
3. headline text : Text of the Headline in English, only ascii characters.

ANALYSIS:

In News Classification and Recommendation System, performance metrics such as Recall, Precision, Accuracy and Mean F-1 score are used to evaluate the performance of the prediction model. These metrics can be used to compare the performance of Naive Bayes Classifier algorithms on the same dataset[7].

1. Recall
   = Ratio of correctly predicted positive observations to the all observations in actual class
   = TP/ [TP + FN]
2. Precision
   = Ratio of correctly predicted positive observations to the total predicted positive observations.
   = TP / [TP + FP]
3. Accuracy
   = Number of correct Predictions
   = [TP + TN] / [TP + FN + FP + TN]
4. F1-Score
   = It is weighted average of precision and recall
   = 2*[Recall * Precision]/[Recall + Precision]

Results of Comparative Analysis of different ML Model

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	86.15	86	86	86
Random Forest	86.15	86	86	86
Multinomial Naive Bayes	92.09	92	92	92
Support Vector Classifier	79.64	80	80	80
Decision Tree Classifier	81.88	82	82	82
K Nearest Neighbor	73.60	74	74	74
Gaussian Naive Bayes	76.09	76	76	76

In the above table, There is the percentage values of the performance parameters of different Machine Learning models. Here, We performed the different Machine Learning algorithms like Logistics Regression, Random Forest, Support Vector Classifier, Decision Tree Classifier, K-Nearest Neighbor, Gaussian Naive Bayes and Multinomial Naive Bayes to find the best algorithm which provides the maximum accuracy. after implementing the Machine Learning algorithms we find the different performance parameters like accuracy, precision, recall and F1-score. After comparing the performance parameters of different ML Model Multinomial Naive Bayes Model gave the maximum accuracy of 92 percentage. So, we used the Multinomial Naive Bayes Model to train the dataset and to get the categories of real time news.

VII. CONCLUSION

Considering the need to classify and recommend the news accurately according to the user age group and its area of interest, naive bayes model provides more accuracy than any other machine learning model. The database used in this system includes the news that were collected from early start of year 2000 to the first quarter of 2022. Dataset contains tremendous amount of news of different categories which increases the scope of news which user reads. The dataset will be used to train the model to classify and recommend the news. The goal of the system for classifying and recommending news will be to provide best and reliable information with high accuracy and a lower error rate. This system will provide news according to its age group and will blocks news which carries the sensitive content. Ultimate goal of the system is to provide accurate and reliable news to the user from the different sources. After implementing and comparing the accuracy of all Machine Learning model, Naive Bayes Model was providing the highest accuracy of 92 percent. Also system is recommending the news to user according to user interests and blocks the sensitive data which is not suitable to user according to user's age.

REFERENCES

[1] Chong Feng, Muzammil Khan, Arif Ur Rahman And Arshad Ahmad, "News Recommendation Systems Accomplishments, Challenges and Future Directions", ResearchGate, vol. 8, pp. 16702-16725, Jan. 2020.

[2] M. Agrawal, M. Karimzadehgan and C. Zhai, "An online news recommender system for social networks", Urbana, vol. 51, pp. 61801, 2009.

[3] M. An, F. Wu, C. Wu, K. Zhang, Z. Liu and X. Xie, "Neural news recommendation with long- and short-term user representations", Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, pp. 336-345, 2019.

[4] S. Athalye, "Recommendation system for news reader", 2013.

[5] I. Rish. An empirical study of the Naive Bayes classifier. IJCAI Workshop on Empirical Methods in Artificial Intelligence, 2001, pp. 825-836.

[6] C. K. Hsieh, L. Yang, H. Wei, M. Naaman, and D. Estrin, Immersive recommendation: News and event recommendations using personal digital traces, in Proc. 25th Int. Conf. World Wide Web (WWW), 2016, pp. 51-62.

[7] N. X. Bach, N. D. Hai and T. M. Phuong, "Personalized recommendation of stories for commenting in forum-based social media", Inf. Sci., vol. 352, pp. 48-60, Jul. 2016.