# News article classification using Naive Bayes Algorithm

**Prasad Rathod[1], Swapnil Gawali[2], Shivprasad Kavathe[3], Amit Dolas[4]**

[1,2,3,4] *Student, Department of Artificial intelligence and data science, Vishwakarma institute of technology, pune, India*

---

**Abstract** – The Naive Bayes classifier, a probabilistic machine learning technique, is useful for classification tasks. It is based on the Bayes theorem, which states that the likelihood of an event occurring given some observed evidence is equal to the prior probability of the event occurring. The Naive Bayes classifier can be trained on a dataset of labelled news articles, each of which is associated with a particular class or category, for the purpose of classifying news articles. The features of the articles, such as the words used and the length of the article, can then be used by the classifier to predict the class of an unseen article. The "naive" assumption, which is one of the key assumptions of the Naive Bayes classifier, is that the articles' features are independent of one another. The classifier is able to predict outcomes without taking into account how features interact with one another because of this assumption. The Naive Bayes classifier can still perform well on many classification tasks, including the classification of news articles, despite this assumption.

***Key Words***: Natural language toolkit ,python ,machine learning algorithm.

## 1. INTRODUCTION

The process of classifying a news article according to its content is known as news article classification. This is a common issue in information retrieval and natural language processing, and it can be useful for organizing and searching through large collections of news articles.

One approach to categorising news stories is to use a machine learning algorithm like the Naive Bayes classifier. A probabilistic model known as the Naive Bayes classifier makes predictions based on the likelihood that particular occurrences will occur. The events are the classes or categories to which news articles can be classified, and the features are the words or other characteristics of the articles.

A dataset of labelled news articles, each of which is associated with a distinct class, is required to train a Naive Bayes classifier for news article classification. The classifier would then learn the probability distribution of the characteristics for each class. which would then use this information to predict articles that had not been seen before. The Naive Bayes classifier is able to efficiently simplify calculations and make predictions because it assumes that the articles' features are independent of one another.

A lot of classification tasks benefit from the Naive Bayes classifier's relative simplicity and ease of use, which is one of its advantages. It can also do well on a variety of classification problems, such as classifying news articles. However, in order to ensure that the classifier is effective, it is essential to evaluate its performance on your particular dataset and problem.

## 2. LITERATURE REVIEW

R. Siva Subhramanian and D. Prabha [22] contributed their paper in In February 2020 on research of This research seeks to identify potential customers.

They used the SBC method to modify the NB model with the goal of enhancing prediction by removing unnecessary dataset features.

According to the experimental findings, the WSNB running time is 0.03 seconds for WSNB at depth 1, 0.06 seconds for WSNB at depth 2, and 0.15 seconds for WSNB at depth 3. Running time for Standard Naive Bayes was 0.16 seconds. Which was unmistakably demonstrating that WSNB shortens the model's running time as compared to traditional Naive Bayes.

Faculty of Agriculture, University of Novi Sad [23] published their article in 2022. The effectiveness of the Naive Bayes approach for predicting water quality was studied by the author. Nine water quality factors were examined, including temperature, oxygen saturation values, and others. Five locations and 68 samples of data were used to assess the water quality using the Naive Bayes model. The testing report ranked each parameter as very good, excellent, good, or bad; after analysing the report and using the method, the author came to the conclusion that the model correctly identified water class in 64 out of 68 instances.

Disha Sharma and Sumit Chaudhary [24] They studied various sources of stress which includes 1) The surrounding Environment 2) Social Stress 3) Physiological 4) Thoughts

Authors applied four machine learning technics that are logistic Regression, Naïve Bayes, Multilayer perceptron ,Bayer's Net.

Parameters like False Positive rate, True Positive Rate , precision, Recall considered for the performance. After comparing all the results of four methods they concluded that Baye's Net classifiers gives longest accuracy of 88 percentage and Naive Bayes gives accuracy of 86 percentage.

Mamata Thakur and team [25] by concerning the problem of huge growth of internet and difficulty in getting relevant topic according to search. Authors chose some news websites after that the important attributes from these

The Nave Bayes algorithm was used by the authors to classify data from 10 different websites, and the results of comparative studies with other current algorithms on the same dataset demonstrate that Nave Bayes outperforms them.

Yi Ying [26] The author of this study employed a variety of news stories to research and used news categories including sports, politics, business, etc. The Confusion Matrix results show that the Sarcasm model developed using the Naive Bayes approach above achieved an accuracy level of 66%, 70% withdrawal, and 68% precision.

By summarising the literature review we can understand sometimes Navies Bayes gives good results but not able to give 100% correct results and some of other machine learning algorithms are more effective than NB, so more researches can be done increase efficiency of NB

## 3. METHODOLOGY

### 3.1 PROPOSED SYSTEM

Here is a general approach to utilising the provided code to categorise news articles:

1. Assemble and classify a dataset of news stories, each with a category tagged (e.g., sports, tech, business, entertainment). The classifier will be trained and tested using this dataset.

2. Remove all stop words from the data and lowercase each word in each article as part of the pre-processing.

3. To turn the text input into numerical feature vectors, create a TfidfVectorizer.

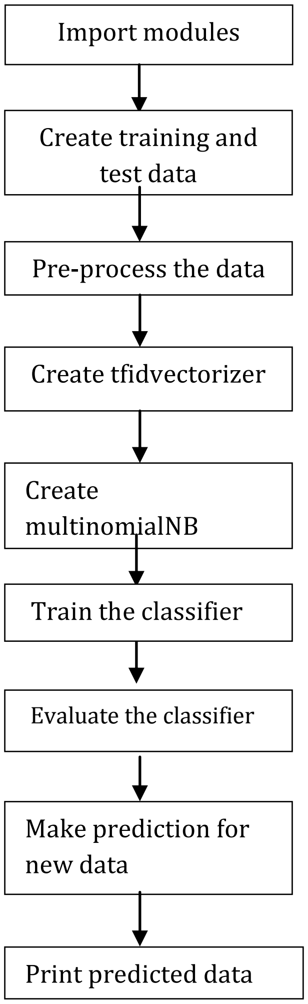
4. The training and test data should be converted into feature vectors using the TfidfVectorizer.

5. Making use of the training data, create a Multinomial Naive Bayes classifier.

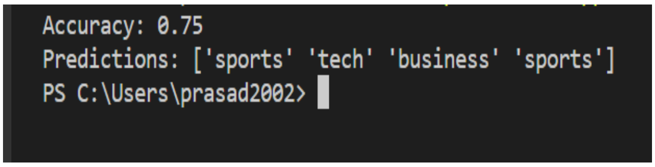
6. Calculate the classifier's accuracy by evaluating it against the test data.

7. Make predictions for fresh, unlabelled news articles using the classifier by converting them into feature vectors and passing them into the classifiers predict method

### a. FLOW CHART



## RESULT AND DISCUSION



```
Accuracy: 0.75
Predictions: ['sports' 'tech' 'business' 'sports']
PS C:\Users\prasad2002>
```

The program we did is a simple classification program that uses naïve bayes classifier to predict the category of text data. The programme first generates two sets of text data: a training set and a test set, where each text is assigned to the appropriate category. The text data is then lowercased and stop words are removed as part of the pre-processing. The text data is then transformed into feature vectors via a Tf-Idf vectorizer, which are then fed into the classifier as input. The classifier is then trained on the training set of data and used to the test set of data to provide predictions.

Because of the limited data set and small amount of data, this programme cannot make good predictions based on the input data.

Use a larger and more varied training and test data set to enhance the effectiveness of this programme. Additionally, you can experiment with various classifiers, feature extraction methods, and text pre-processing approaches. It can also be beneficial to fine-tune the model parameters, use pre-trained models, and fine-tune it using your own dataset.

Before selecting a particular technique or model, it's crucial to take the context of the problem at hand and the precise requirements of the task into account.

## 3. CONCLUSION

The Naive Bayes classifier is a popular machine learning method that can be used to categorise news stories. It has a lot going for it, like being easy to use, working well, and doing well on a lot of classification tasks. However, its performance on a particular dataset must be evaluated. The classifier's performance can be improved, more complex problems can be handled, and the classifier can be applied to new domains can all be developed further in this area. Natural language processing and information retrieval could benefit greatly from using the Naive Bayes classifier.

## FUTURE SCOPE:

In the field of news article classification using Naive Bayes classifiers, there are numerous potential future directions for research and development. These are some: expanding the application of the classifier to new domains and languages, enhancing the classifier's performance, and incorporating it into news analysis systems. The Naive Bayes classifier offers a lot of potential for solving a variety of real-world issues, and more research may be done on its capabilities and restrictions when it comes to categorising news items.

## REFERENCES

[1] Kuldeep Vayadande, Aditya Bodhankar, Ajinkya Mahajan, Diksha Prasad, Shivani Mahajan, Aishwarya Pujari and Riya Dhakalkar, "Classification of Depression on social media using Distant Supervision", ITM Web Conf. Volume 50, 2022.

[2] Kuldeep Vayadande, Rahebar Shaikh, Suraj Rothe, Sangam Patil, Tanuj Baware and Sameer Naik," Blockchain-Based Land Record System", ITM Web Conf. Volume 50, 2022.

[3] Kuldeep Vayadande, Kirti Agarwal, Aadesh Kabra, Ketan Gangwal and Atharv Kinage," Cryptography using Automata Theory", ITM Web Conf. Volume 50, 2022

[4] Samruddhi Mumbare, Kunal Shivam, Priyanka Lokhande, Samruddhi Zaware, Varad Deshpande and Kuldeep Vayadande,"Software Controller using Hand Gestures", ITM Web Conf. Volume 50, 2022

[5] Preetham, H. D., and Kuldeep Baban Vayadande. "Online Crime Reporting System Using Python Django."

[6] Vayadande, Kuldeep B., et al. "Simulation and Testing of Deterministic Finite Automata Machine." International Journal of Computer Sciences and Engineering 10.1 (2022): 13-17.

[7] Vayadande, Kuldeep, et al. "Modulo Calculator Using Tkinter Library." EasyChair Preprint 7578 (2022).

[8] VAYADANDE, KULDEEP. "Simulating Derivations of Context-Free Grammar." (2022).

[9] Vayadande, Kuldeep, Ram Mandhana, Kaustubh Paralkar, Dhananjay Pawal, Siddhant Deshpande, and Vishal Sonkusale. "Pattern Matching in File System." International Journal of Computer Applications 975: 8887.

[10] Vayadande, Kuldeep, Ritesh Pokarne, Mahalakshmi Phaldesai, Tanushri Bhuruk, Tanmay Patil, and Prachi Kumar. "Simulation Of Conway's Game of Life Using Cellular Automata." SIMULATION 9, no. 01 (2022).

[11] Gaurav, Rohit, Sakshi Suryakant, Parth Narkhede, Sankalp Patil, Sejal Hukare, and Kuldeep Vayadande. "Universal Turing machine simulator." International Journal of Advance Research, Ideas and Innovations in Technology, ISSN (2022).

[12] Vayadande, Kuldeep B., Parth Sheth, Arvind Shelke, Vaishnavi Patil, Srushti Shevate, and Chinmayee Sawakare. "Simulation and Testing of Deterministic Finite Automata Machine." International Journal of Computer Sciences and Engineering 10, no. 1 (2022): 13-17.

[13] Vayadande, Kuldeep, Ram Mandhana, Kaustubh Paralkar, Dhananjay Pawal, Siddhant Deshpande, and Vishal Sonkusale. "Pattern Matching in File System." International Journal of Computer Applications 975: 8887.

[14] Vayadande, Kuldeep B., and Surendra Yadav. "A Review paper on Detection of Moving Object in Dynamic Background." International Journal of Computer Sciences and Engineering 6, no. 9 (2018): 877-880.

[15] Vayadande, Kuldeep, Neha Bhavar, Sayee Chauhan, Sushrut Kulkarni, Abhijit Thorat, and Yash Annapure. Spell Checker Model for String Comparison in Automata. No. 7375. EasyChair, 2022.

[16] VayadandeKuldeep, Harshwardhan More, Omkar More, Shubham Mulay, Atharva Pathak, and Vishwam Talnikar. "Pac Man: Game Development using PDA and OOP." (2022).

[17] Preetham, H. D., and Kuldeep Baban Vayadande. "Online Crime Reporting System Using Python Django."

[18] Vayadande, Kuldeep. "Harshwardhan More, Omkar More, Shubham Mulay, Atahrv Pathak, Vishwam Talanikar,"Pac Man: Game Development using PDA and OOP"." International Research Journal of Engineering and Technology (IRJET), e-ISSN (2022): 2395-0056.

[19] Ingale, Varad, Kuldeep Vayadande, Vivek Verma, Abhishek Yeole, Sahil Zawar, and Zoya Jamadar. "Lexical analyzer using DFA." International Journal of Advance Research, Ideas and Innovations in Technology, www. IJARIIT. com.

[20] Manjramkar, Devang, Adwait Gharpure, Aayush Gore, Ishan Gujarathi, and Dhananjay Deore. "A Review Paper on Document text search based on nondeterministic automata." (2022).

[21] Chandra, Arunav, Aashay Bongulwar, Aayush Jadhav, Rishikesh Ahire, Amogh Dumbre, Sumaan Ali, Anveshika Kamble, Rohit Arole, Bijin Jiby, and Sukhpreet Bhatti. Survey on Randomly Generating English Sentences. No. 7655. EasyChair, 2022.

[22] R.Siva Subramaniyam ,D.Prabha Customer behavior analysis using weighted naïve Bayesian background "international journal of innovative technology and exploring engineering (IJITEE)

[23] Department of water management faculty of agriculture ,university of Novi sad, Water quality prediction based on naive Bayes algorithm(2022)

[24] Disha Sharma, Sumit Chaudhary "stress prediction of professional student using machine learning" .international journal of engineering and advanced technology (IJEAT)

[25] Mamata thakur ,Priyanka thakur ,Pritam thakur ,govinda rao meetu."Classification of news using naïve algorithm". International journal of creative research thoughts (IJCRT)(2018).

[26] Yi Ying ."eff Effectiveness of the News Text Classification Test Using the Naïve Bayes". Journal of physics: conference series (2021).

## BIOGRAPHIES

Praasad Namdeo Rathod .
"Student, Department of Artificial intelligence and data science, Vishwakarma institute of technology, pune, India "

Swapnil Lahu Gawali.
""Student, Department of Artificial intelligence and data science, Vishwakarma institute of technology, pune, India "
"

Shivprasad Vyankat Kavathe.
""Student, Department of Artificial intelligence and data science, Vishwakarma institute of technology, pune, India "
"

Amit Vasant Dolas.
""Student, Department of Artificial intelligence and data science, Vishwakarma institute of technology, pune, India "
"