Fake News Detection using Naïve Bayes and Support Vector Machine

1st Vasireddy Sneha
dept. of Computer Science
B V Raju Institute of Technology
Narsapur, India
21211a05x6@bvrit.ac.in

2nd Velaga Sravani
dept. of Computer Science
B V Raju Institute of Technology
Narsapur, India
21211a05y3@bvrit.ac.in

3rd Vedantham Sathvika
dept. of Computer Science
B V Raju Institute of Technology
Narsapur, India
21211a05x8@bvrit.ac.in

4th Banoth Venu Kumar
dept. of Computer Science
B V Raju Institute of Technology
Narsapur, India
22215a0532@bvrit.ac.in

5th Dr.A.Jagan
dept. of Computer Science
B V Raju Institute of Technology
Narsapur, India
jagan.amgoth@bvrit.ac.in

***Abstract*—Social media platforms have revolutionised how people engage with the outside world by allowing people to voice their thoughts and exchange information on a variety of topics that interest them. However, because social media is so widely used, information spreads quickly among thousands of users, giving it the perfect environment for the fast spread of false information. The alarming rise of false news presents major concerns to both users and the nation, demanding prompt action. On a national and personal level, the potential harm that misinformation can do calls for serious thought and preventive action. Online fake news identification has garnered substantial research interest, yet the results of past endeavors, particularly using the naive Bayes classifier, have yielded suboptimal performance. In response, this work aims to go deeper into the topic and address the constraints of the naive Bayes classifier by investigating techniques to improve its efficiency. To efficiently identify fake news on social media, we provide a machine learning-based technique that makes use of Naïve Bayes and Support Vector Machine classifiers. Our main goal is to provide users with a tool that enables them to separate significant information from false information and determine the reliability of news stories they come across online.**

***Index Terms*—Fake news, authenticity, Navie bayes, Support Vector Machine (SVM), Count vectorizer, Term Frequency-Inverted Term Frequency (TF-IDF), Natural Language Processing (NLP), Natural Language Tool Kit (NLKT)**

I. Introduction

Due to increasing amount of time we spend communicating online through social media platforms, many people are discovering and reading news there. Social media is currently a more important news source than any other media. Despite the benefits, social media stories are of poorer quality than those published by conventional news outlets. Because it is inexpensive to provide news online and the news travels more rapidly and readily through social media, a lot of false news that is, news articles containing purposely inaccurate information is published online for many reasons, including financial and political profit. False information effects how individuals view and respond to actual news. Some false news is simply intended to stimulate people's suspicions and doubts. Fake news can be widely disseminated and cause major damage for people and society. Fake news deliberately encourages people to adopt prejudiced or false beliefs. False news is a common tool used by propagandists to spread ideologies or gain power. Confusion and issues arise when it becomes harder to discern between reliable and misleading news. Manually identifying fake news is difficult; it can only be done when the person doing the identification has a deep expertise of the news topic. There is a need to develop techniques for identifying fake news stories on social media. To prevent false rumours from propagating on social media and messaging applications, a fake news detection technology is used. Identifying false news and authentic news is the aim of fake news detection. The approach outlined in this work may be used to develop a model that evaluates the veracity of news based on its origins, terms, and phrases. To implement this concept into execution, supervised machine learning methods can be employed. Then, using the confusion matrix as a guide, feature selection methods are implemented to experiment and choose the accurate features. We recommend applying various categorization method to build the model. The resultant model will include testing the unobserved data. The end result will be a model that recognises and categorise fake news and articles. Any system can use this model and incorporate it for later usage.

II. Literature Survey

Many automatic detection strategies for fake news have been documented in the literature, with the majority of them focusing on identifying internet available social media posts. Because the database used contained text, different types of Natural Language Processing models are used to transform the data into the desired format.

In a realistic strategy [12] for detecting fake news on Twitter was developed, in which large scale datasets from Twitter are collected for training a machine learning model, as well as a few small, hand-labelled datasets for evaluation reasons. The following features were considered: text, topic, and sentiment. Learning approaches included naive bayes, decision trees, support vector machines, and neural networks. The results demonstrated that using both tweet and source features generated better results than using only tweet features.

It was suggested to use n-gram analysis and machine learning approaches in a false news detection model [14]. Six different classification algorithms and two separate feature extraction procedures were contrasted in this model. The results of the studies reveal that the method TF-IDF yields the best results. They employed the Linear Support Vector Machine (LSVM) classifier, which has a 92% accuracy.

A straight forward method [13] for detecting fake news using a naive Bayes classifier was provided. This method is being evaluated using data taken from social media news posts. It is stated that the model may attain 74% accuracy. The performance of a classifier may be improved by training it on a larger dataset with more extensive news articles, eliminating stop words, using stemming, treating uncommon terms differently, and computing probabilities using groupings of words rather than single words. Although this model's rate is respectable, several other articles have achieved greater rates by using various classifiers. However, this study proved that even a simple classifier can get good results on such classification problems.

The authors of [10] evaluated the performance of several approaches using three different data sets. This work ignores elements such the source, author, and publication date, which may have a substantial impact on the outcome and instead concentrates on the content and emotion of the material. Few studies treated each tweet or post as a challenge for binary categorization. The original location of the post was the only factor considered for categorising. The Twitter API was used by the authors to acquire the datasets. The following algorithms were applied to data sets: 1. Bayes 2. Decision Trees 3. A support vector device 4. Convolutional neural networks 5. Random forest 6. XG Boost. The results revealed that 15% of the tweets were phony, 45% were authentic, and the remaining messages were undetermined.

The methodology proposed in [11] for automatically detecting fake news in internet news used the datasets, that were obtained straight from the internet and through an approach of manual data collection and internet assistance. Various analyses were done on these datasets to discover linguistic features that are commonly found in untrue content. The accuracy of creating a fake news detector based on linguistic variables was up to 78%.

A Random forest based model [9] was developed to detect fraudulent news using five characteristics: amusement, adverse effects, illogicality, syntax, and symbols. Its purpose was assessing the veracity of a stories. Researchers began by obtaining the web address of the article that has to be judged as true or false. The content of web address is then collected. The content is then routed to multiple data pre-processing elements. Their model has an accuracy rate of 87%.

An ensemble learning approach [7] was designed in which multiple base classifier such as SVM, Convoluted neural networks, Long short term memory, KNN and naïve bayes classifier is used. The background of news is reviewed, and a trustworthiness rating for the news and the origin of the claim is generated. Feature extraction approaches such as bag of words, n-grams method, and TF-IDF were employed in this model. Linguistic traits are very useful for detecting deceptive material. Naive Bayes, CNN, and LSTM worked together to achieve 94% accuracy. An average accuracy rate of 85% was found.

Another method [5] was developed to extract a collection of elements that can identify untrue news from a collection of news that has been pre-processed using purifying methods, stemming, N-gram method, Term frequency-Inverted document frequency, bag of words. The classifier used is a SVM classifier. It was shown that the "Sentiment" feature had almost no impact on precision, which seems rational given that just because a comment elicited a negative reaction does not necessarily mean that it was fabricated. The author column increased the accuracy to 100%, demonstrating the efficiency of the recommended encoding. The following are the best indicators of fake news are text, writer, origin, date, and emotion. With large texts and extensive datasets, the N-gram approach outperforms the bag of words.

Natural language processing technique was implemented [8] in which Term frequency document frequency (TF-IDF) was employed for extraction of characteristics. A word's importance increases in direct proportion to how often it appears in the manuscript. Classifiers used include Naive Bayes, Support Vector Machine, and passive aggressive. Python Natural language toolkit (NLTK) is used to implement these classifiers. On the dataset, the passive aggressive classifier had an accuracy of 92.9%, the naive bayes classifier had an accuracy of 84.056%, and the support vector machine method attained 95.05%. The proposed model achieves the highest precision with the SVM classifier.

A method [6] was proposed that gathers crucial characteristics from datasets of false news and categorises them using a combination model made up of three widely used machine learning approaches. An ensemble technique is a method for combining the findings of different machine learning algorithms to get results that are more accurate. For the proposed strategy, well-known false news datasets like liar and

ISOT were located and examined. This model's training and testing accuracy on the ISOT dataset were 99.8 percent and 44.15 percent, respectively. On the other dataset, it achieved complete validation and training efficiency. Performance was enhanced by carefully choosing the features to use and by adjusting the hyper parameters.

III. Existing Work

In the approach proposed in [9], the dataset was erratically mixed and then partitioned into subsets. The learning set was used to train naive Bayes classifier. Some global classifier features were adjusted using the validation dataset. On the validation dataset, the absolute likelihood of the fact produced the best results. True and false news pieces have nearly the same categorization accuracy, however fake news has significantly lower classification accuracy. This could be related to dataset skewness.

The authors of [1] claim that a dataset made by fusing of real news with a fake news was used to assess the system's performance. Then, each feature is put to the test in turn until the highest accuracy rate is attained. They were able to gather a dataset that comprised the following components: publication date, emotion, source, author, and category. Additionally, they collected five terms using the bag of words strategy and three words using the N-gram method. The support vector machine (SVM) algorithm was employed. Text, author, source, and date are the top features for detecting fake news. With large texts and extensive datasets, the N-gram approach outperforms the bag of words. Sentiment analysis produced very poor accuracy.

IV. Proposed Work

The proposed work intends to use count vectorization method and TD-IDF for feature extraction and two machine learning classifiers Naive bayes and Support vector machine to train the model. For system training and testing, a sizable dataset of actual and fraudulent news articles is gathered. The data must go through several modifications,

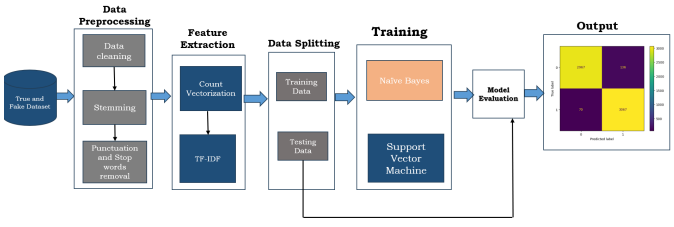

Fig. 1. Architecture diagram of proposed model

including stop words removal, tokenization, lower casing, sentence segmentation, and punctuation removal. By deleting any unnecessary data from the data, we can decrease the amount of the actual data. Stop words are inconsequential words that make noise when utilised as characteristics in text categorization. After tokenization, the tokens must then be converted into a standard form. After the data has been tokenized, Stemming will then be carried out. Stemming is the straightforward process of returning words to their original form while lowering the number of word types or classes in the data. In order to extract features, TF-IDF and count vectorization is employed. The Term Frequency-Inverted Document Frequency (TF-IDF) is a weighting metric that is commonly used in feature extraction and NLP. Training the classifier is the final step in the classification process. The dataset is divided into training and testing sets. Testing set is used to validate and training set is used to train the classifier. For classification, the support vector machine and naive bayes algorithms are employed. Naive Bayes is a popular classification technique for detecting fake news. The core thought behind Naive Bayes is to utilise Baye's theorem to assess the likelihood of a particular piece of news being false based on the probabilities of noticing certain elements in the news. SVMs work by locating a hyperplane that divides data into two classes, in this instance, legitimate and fraudulent news. It can be used to classify new news articles as legitimate or fraudulent by measuring the distance between the feature vector of each post and the hyperplane. Training ends if the model is accepted and can be utilised. The learning algorithm's parameters are updated to if the model is not up to standard.

V. Methods

A. Dataset description

A dataset from Kaggle known as the fake news dataset is chosen for this work. Information about news reports and their labels are included in the dataset. There are five characteristics in the dataset: "Id," "Title," "Text," "Author," and "Label". The dataset doesn't disclose any details on the precise domain or source of the news articles. The dataset contains 20800 records, 10387 of which are considered fake news, while the remaining 10413 are deemed true news. The whole data set is available in a CSV (comma-separated values) format, which is read and transformed into a data frame using the Pandas module of Python. After being normalised with the Python sklearn package, the dataset is divided into training and testing sets. 30% of the entire dataset is considered as the testing set.

TABLE I
Dataset Schema

Sno	Column	Dtype
1	Id	int64
2	Title	object
3	Author	object
4	Text	object
5	Label	int64

B. Data pre-processing

Certain modifications must be applied to the dataset to minimize the amount of the data by deleting unnecessary data. Punctuations have no impact on text classification. Stop words are meaningless words that have no role when used

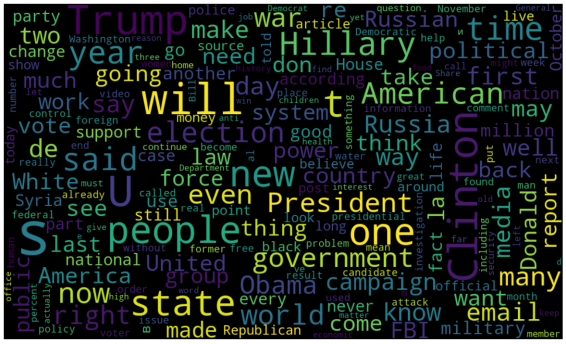

Fig. 2. Word cloud of true news.

Fig. 3. Word cloud of fake news.

as text categorization criteria. Stop words are regarded to be articles, prepositions, conjunctions, and pronouns. Stemming is the straightforward process of returning words to their native form while lowering the number of word types. For example, the words “playing”, “player”, “played” will be reduced to the word “play”. The data must thus undergo a variety of adjustments, such as removal of punctuation, stop words, and then stemming.

C. Feature extraction

The main aim of this method is to select new features from the present ones in a dataset, hence reducing the overall number of features in the dataset. This new reduced collection of features should then be able to put together the majority of the information in the original set of features. The proposed work intends to use n-grams method and TD-IDF for extraction of features.

Count Vectorizer analyses how many times each word appears in a article in order to convert text into vectors of term or token counts. In the matrix generated through the Count Vectorizer, each distinct word is represented by a column, and each sample of text from the article is represented by a row. The value of each cell is just how many times a word occurs in that particular text sample.

The Term Frequency-Inverted Document Frequency (TF-IDF) is a commonly employed metric in feature extraction and Natural Language Processing (NLP). The more times a phrase occurs in a text, the more significant it becomes, but the frequency of the term balances this out.

D. Dataset Splitting

Following the feature extraction phase, the dataset was partitioned into two distinct subsets to facilitate the machine learning process. 70% of the data was designated as the training set, which served as the basis for developing and fine-tuning the machine learning models. The testing set, which was given the remaining 30% of the dataset, was crucial in evaluating these models’ performance in an objective manner. This explicit separation of the data into training and testing allowed for the construction of strong models and a valid assessment of their ability to identify false news stories.

E. Learning

Training the classifier is the final step in the classification process. Sets for training and testing are separated from the dataset. Testing set is used to validate and training set is employed to instruct the classifier. For classification, support vector machine and naive bayes algorithms can be applied.

Naive Bayes is a frequently used method for classification problems. The core thought behind Naive Bayes is to utilise Baye’s theorem to assess the likelihood of a particular piece of news being false based on the probabilities of noticing certain elements in the news. Its performance is determined on the quantity and quality of features employed, in addition to the quantity and state of the training dataset.

SVMs work by locating a hyperplane that divides data into two classes, in this instance, legitimate and fraudulent news. The distance between each post’s feature vector and the hyperplane can be used to determine if a news story is authentic or counterfeit. The article is regarded as authentic if the distance is positive; it is regarded as fraudulent if the distance is negative. Working with non-linearly separable data makes SVMs an especially effective and popular method for spotting fake news.

VI. Result

The dataset used in this model is collected from Kaggle. The structure of dataset is (20800, 7). The title and text columns of the dataset are concatenated. For removal of stop words, Natural language tool kit (NLKT) library is employed. Porter Stemmer is implemented, which is a frequently used stemming algorithm because of its accuracy.

After pre-processing, 70% of the dataset is taken take into consideration for training, while the remaining 30% is considered for testing. Count Vectorizer and TF-IDF are applied for feature selection. To gain deeper insights into the model’s performance, we presented the confusion matrices for both the SVM and Multinomial Naïve Bayes classifiers in Figures 4 and 5, respectively. These matrices provide a breakdown of true positives, true negatives, false positives, and

false negatives, allowing us to assess the models' behavior in various scenarios.

- True positive: actual and expected outcomes are positive.
- True Negative: actual and expected outcomes are negative.
- False positive: predicts a positive outcome but is actually negative.
- False negative: predicts a negative outcome but is actually positive.

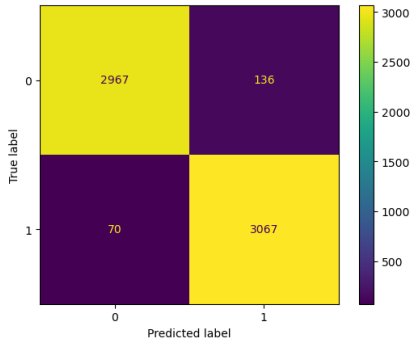

Fig. 4. Confusion matrix obtained using Support Vector machine classifier

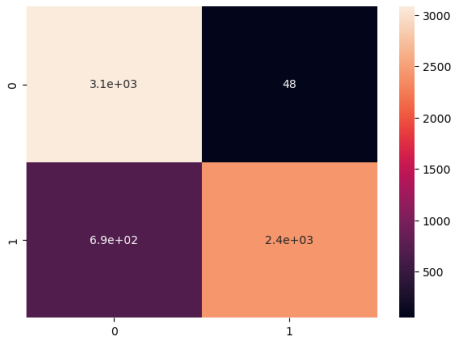

Fig. 5. Confusion matrix obtained using Naive bayes classifier

TABLE II
PERFORMANCE METRICS

Classifier	Recall	Precision	F-Measure	Accuracy
Proposed Multinomial Naïve Bayes	88%	90%	88.98%	88.20%
Proposed Support Vector Machine	97%	96.9%	96.94%	96.69%
Naïve Bayes[13]	13%	71%	21.9%	74%
Support Vector Machine [8]	93.73%	92.56%	93.14%	95.05%

The number of instance positives that genuinely match to the positive instances is referred to as precision. Recall measures how many accurate instance predictions were produced using all the positive instances in the collection. The F-measure demonstrates an equilibrium between accuracy and recall.

$$\text{F-Measure} = \frac{2 \times \text{p} \times \text{r}}{\text{p} + \text{r}} \quad (1)$$

Where,

p indicates precision

r indicates recall

The Multinomial Naïve Bayes algorithm achieved an accuracy rate of 88.20%. It classifies news items with a respectable level of accuracy, but in terms of total performance, it is outperformed by the Support Vector Machine (SVM). SVM excels in recall, precision, and F-Measure while attaining a high accuracy rate of 96.69%. These measures show that SVM is a reliable option for identifying fake news.

VII. Conclusions

The study offers an machine learning-based method that makes use of elements like title, content, and author information to identify fake news on social media. The performance evaluation demonstrates that the proposed algorithm outperforms both the conventional naive Bayes and support vector machine algorithms. The TF-IDF (Term Frequency-Inverse Document Frequency) approach significantly improved the effectiveness of the naive Bayes classifier. This emphasises the significance of preprocessing and feature engineering in increasing the classifier's capabilities. The experimental results reveal that the proposed model achieves an impressive accuracy rate of 96.69% in detecting fake news, signifying the effectiveness of the developed approach. However, it is crucial to acknowledge that the fight against fake news is a never-ending difficulty since perpetrators constantly modify their strategies to avoid discovery. Therefore, ongoing monitoring, updating, and optimisation of the algorithm will be necessary to preserve it's effectiveness and relevance in a constantly changing online environment.

References

[1] Musunuri Naga Venkata Vinay Babu, Vukoti Vineel Kumar, Turaga Karthik Vedavyas, Veerraju Gampala, Swarna Dinesh Chandra, Satish Thatavarthi, " Machine learning approaches for fake news detection", IEEE, 2023.

[2] Nitish Kumar, Nirmalya Kar, "Approaches towards Fake news detection using machine learning and deep learning" in 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN). IEEE, 2023.

[3] Zhiwei Guo, Keping Yu, Gang Li, Feng Ding, Amin Beheshti, Alireza Jolfaei, "Mixed Graph Neural Network – Based Fake news detection for sustainable vehicular social network", in IEEE Transactions on Intelligent Transportation Systems. IEEE, 2022.

[4] Reshmi T S, Daniel Madan Raja S "Fake news detection using voting Ensemble classifier" in 2022 International Conference on Inventive Computation Technologies (ICICT), IEEE, 2022.

[5] Nihel Fatima Baarir, Abdelhamid Djeffa, " fake news detection using machine learning", IEEE 2021.

[6] Saqib Hakak , Mamoun Alazab , Suleman Khan, Thippa Reddy Gadekallu, , Praveen Kumar Reddy Maddikunta , Wazir Zada Khan, " An ensemble machine learning approach through effective feature extraction to classify fake news", 0167-739X/© 2020 Elsevier.

[7] Arush Agarwal, Akhil Dixit, "Fake News Detection: An Ensemble Learning Approach", in Proceedings of the International Conference on Intelligent Computing and Control Systems. IEEE, 2020.

[8] Jasmine Shaikh, Rupali Patil, "Fake news detection using machine learning", in International Symposium on Sustainable Energy, Signal Processing and Cyber Security (ISSSC). IEEE, 2020.

[9] Manisha Gahirwal, Sanjana Moghe, Tanvi Kulkarni, Devansh Khakhar, Jayesh Bhatia, "Fake news detection", International Journal of Advance Research, Ideas and Innovations in Technology ISSN, 2019.

[10] Syed Ishfaq Manzoor, Dr Jimmy Singla, Nikita "Fake News Detection Using Machine Learning approaches: A systematic Review", in Proceedings of the Third International Conference on Trends in Electronics and Informatics. IEEE, 2019.

[11] Pérez-Rosas, Verónica Kleinberg, Bennett Lefevre, Alexandra Mihalcea, Rada, "Automatic Detection of Fake News", in proceedings of 27th International Conference on Computational Linguistic. Association for Computational Linguistics, 2018.

[12] S. Helmstetter and H. Paulheim, "Weakly supervised learning for fake news detection on Twitter," Proc. 2018 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2018, pp. 274–277, 2018.

[13] Mykhailo Granik, Volodymyr Mesyura, "Fake News Detection Using Naive Bayes Classifier", First Ukraine Conference on Electrical and Computer Engineering (UKRCON). IEEE, 2017.

[14] Hadeer Ahmed, Issa Traore, and Sherif Saad, "Detection of online fake news using n-gram analysis and machine learning techniques", in International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, pages 127–138. Springer, 2017.