IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

FAKE NEWS CLASSIFIER

[1]Aviral Bhardwaj, [2]Kartikay Naswa, [3]Sachin Adlakha

[1]Designation of 1st Author, [2]Designation of 2nd Author, [3]Designation of 3rd Author
[1]Name of Department of 1st Author,
[1]Name of organization of 1st Author, City, Country

1 Abstract

False information that harms society by deceiving people's perceptions and views has a negative impact on the credibility of news that is actually reported. In society, it behaves like a virus. Due to unmoderated channels like personal social media accounts, blog sites, etc., fake news is disseminating more widely. Therefore, it is essential to find it and prevent it from spreading.

The study that was used to create this paper consists mostly of three stages: the first is data pre-processing, which comprises data cleaning and stopword removal. the analysis of our dataset utilizing natural language processing (NLP) for testing. We used machine learning techniques to train the model and obtain predictions from the model after applying NLP. Python has been used with a variety of tools. Pandas for data manipulation and analysis, NLTK for stop word removal and lemmatization, Scikit Learn for our machine learning models, as well as seaborn and matplotlib for data visualization and output. By carefully choosing features, effectively fine-tuning parameters, and utilizing a well-balanced dataset from Kaggle, accuracy of 94.5% has been attained in this paper.

2 INTRODUCTION

False information is referred to as fake news, and it is a serious problem that needs to be resolved. There has always been fake news. People have abused it to further their own political and/or personal objectives. After the printing press was created in the fifteenth century, news began to spread more quickly. Real news spreads just as quickly as fake news. Real and fake news spread like wildfires as a result of improved technology, the internet, and social media. The results could be severe if they are not identified and stopped in time. As a result of rising demand for immediate news updates, news distribution platforms have evolved. Newspapers that publish news a day later have made way for social media updates, app notifications from news applications, and forwarded chat group communications. With such a large number of sources, false information can readily be spread alongside legitimate news. This might become an issue if people begin to believe everything they read in the news.

While organizations like the Press Information Bureau keep an eye out for false information posted online, manually identifying fake news is time-consuming and ineffective. With thousands of stories being published every day, it becomes impossible to carefully review every single piece. Because of this, it is now necessary to develop an automated fake news detector that uses current technology in order to lessen the inconvenience of doing so.

Using supervised machine learning techniques on a labeled dataset, this paper suggests a method for building a model that can assess the reliability of a news story based on its headline, words, and phrases. Here, stopwords are eliminated, lemmatization is performed, and data is cleaned using NLP. To determine whether the news is fake or not, we use supervised learning techniques in the model and classification models within supervised learning. And finally, use the confusion matrix and evaluation metrics to evaluate the test data in order to acquire the best result from the model.

DATASET

Since a dataset is the first prerequisite for a machine learning model, we got our data from Kaggle and cleaned it to meet our specifications. Here are the cleaning techniques.

3 DATA PREPROCESSING

The dataset is cleaned and pre-processed in this stage before training the model. Here, the NLTK library is used. It is a well-known NLP library [3]. It has all the built-in features we require to simplify our job, including stopword removal, punctuation removal, tokenization, and lemmatization of the dataset to improve accuracy and fit the model.

Preprocessing the text is what we are doing in this case because it is the first step in every NLP project. Preprocessing incoming text entails converting the data into a format that is predictable and easy to analyze. It's an important step in developing an outstanding NLP application.

3.1 TOKENIZATION

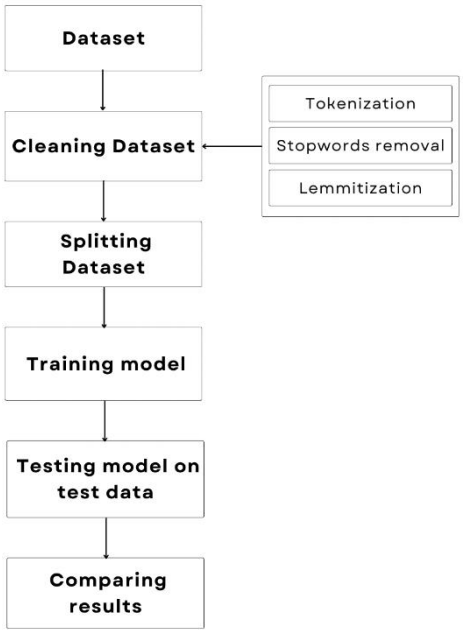
Tokenization is the process of turning a string into a numerical data structure suitable for machine learning. Tokenization is considered as the first step in any NLP pipeline.

3.2 STOPWORD REMOVAL

Stopwords must be eliminated after the text has been tokenized. Stopwords like "the," "by," and other similar words are discovered to be useless in a string and have no special meaning in the text, therefore eliminating them is another crucial step. Therefore, we should eliminate terms like these because they don't help train the model and may lead to issues with it.

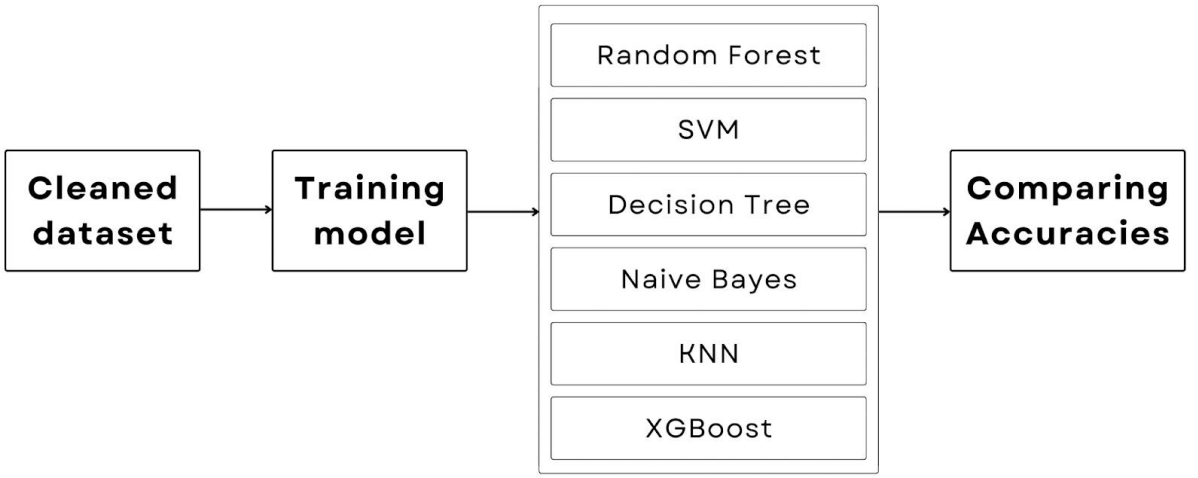
3.3 LEMMATIZATION

Natural Language Processing employs the text normalization method known as lemmatization (NLP) . The issue of words with similar meanings but different texts is resolved by lemmatization. For example, although study and studying have the same meaning, they are treated differently when not lemmatized. Thus, lemmatization helps us here by merging all the words with the same meaning into a single word.



4 MODELS

In this paper we are using a variety of classification models in supervised learning.



4.1 RANDOM FOREST

Random forest is a decision tree model that works on a bagging-based algorithm. It is one of the most popular and best-suited algorithms in the field of machine learning. Both classification and regression datasets can be processed using random forest, which is a supervised learning technique. Based on decision tree techniques, the random forest algorithm was developed. Decision trees are a common component of random forests, to put it simply. With extremely high accuracy and precision scores, random forest can categorize big datasets with ease. During training, it generates a huge number of decision trees, and iterations of training data influence the outcome of the models projected by each tree. A majority vote can be used to merge decision trees for prediction, or all decision trees can be averaged.

4.2 SVM

Another well-known model in machine learning is the SVM method [10]. Under supervised learning, SVM falls. The SVR model can handle regression issues, whereas it can only handle classification issues. The SVM method divides the dataset into classes using a hyper-plane in n dimensions or a line. The hyperplane's primary function is to divide the classes into other classes so that they might differ from one another. Additionally, it addresses the issue of overfitting, which in some situations can compromise the model's accuracy.

4.3 DECISION TREE

Another often used model in machine learning is the decision tree. The supervised learning branch includes decision trees. It can be applied to both the classifications dataset and the regression dataset. Here, we've employed a classification strategy based on decision trees. A decision tree is a decision-making model that resembles a tree. By giving branches to each of the top nodes in a tree, decision trees can handle high-dimensional data since they base their decisions on the conditions that are followed by the nodes. Test nodes or intermediate nodes are nodes with outgoing edges. Nodes are known as leaves at the lowest levels. To increase accuracy, categorization models frequently use tree-based techniques.

4.4 NAIVE BAYES

Naive Bayes is a classification model that belongs to the area of supervised learning. As a Bayes theorem-based probabilistic classifier, it bases its predictions on the chances that an object will be discovered. Text classification, spam filtration, and sentiment analysis are all frequent applications of the Naive Bayes algorithm. The probability P(y|x) of each class, y, of a certain object, x, is estimated via naive bayes. The estimation can be applied to categorization tasks.

4.5 BAYES THEOREM

A straightforward formula that can be used to determine conditional probabilities is the Bayes' Theorem.
Conditional probability is a measure of the likelihood that an event will occur if a different event has previously occurred (by assumption, presumption, assertion, or evidence).

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

Posterior probability (P(A|B)): is the probability of hypothesis A on the observed event B.
Likelihood probability P(B|A): stands for Likelihood, which is the probability of the evidence provided that a hypothesis' probability is true.
Prior Probability (P(A)): is the probability of a hypothesis before seeing the evidence. Marginal Probability P(B): stands for Probability of Evidence Marginal Probability

4.6 KNN

The supervised learning algorithm K-Nearest Neighbor (KNN) can be applied to classification and regression issues. KNN finally, use new positions using the majority of the sounds from the nearby k in relation to them. According to the role of the Euclidean distance, the assigned position in the class is highly mutually exclusive between the closest neighbors K. The value of K determines how well the entire model functions. The K value plays a crucial role. These are some methods that can be used to calculate K's value

4.7 XGBOOST

XGBoost is a decision-tree-based ensemble Machine Learning technique that employs a gradient boosting framework. Tasks involving classification and regression, among others, can benefit from gradient boosting. It presents an expectation model as a group of mediocre decision trees for forecasting. Decision trees are generated sequentially in this approach. Weights are significant in XGBoost. Each independent variable is given a weight before being fed into the decision tree that forecasts outcomes. The variables are subsequently fed into the second decision tree with an enhanced weight for variables that the tree incorrectly anticipated. These distinct classifiers/predictors are then combined to produce a robust and accurate model. It can be used to solve problems including regression, classification, ranking, and custom prediction.

5 RESULTS

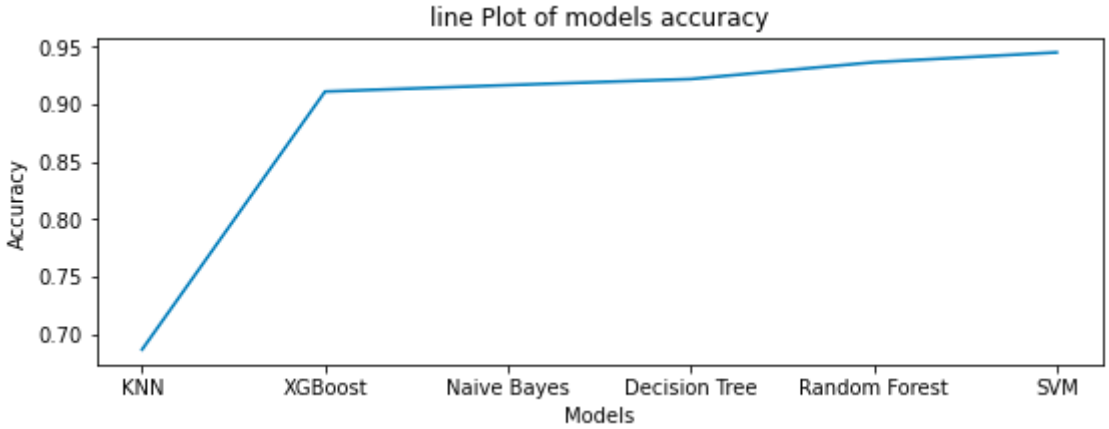
This research paper has trained and tested the dataset on 6 models. The six models are as follows:

- Random Forest Classifier
- Naive Bayes
- SVM
- Decision Tree Classifier
- XGBoost
- KNN

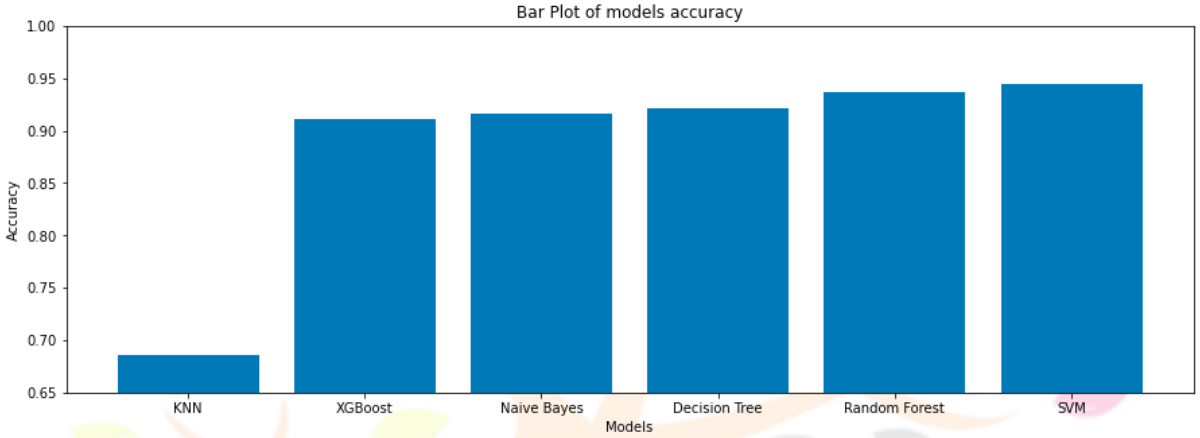
5.1 ACCURACY

The SVM model, out of all the ones tested, had the highest accuracy (94.5%). In the provided dataset, the KNN fared the worst and had an accuracy of 68%.
The accuracy of each model examined is displayed in the graph below.

This is the model's accuracy line plot. The line on the graph starts at 65% and rises to 100%.



This is the accuracy bar plot for the model. The graph begins at 65% and progresses to 100%.



You can conclude how the model has grown by looking at these two graphs.

5.2 Evaluation Metrics

In order to assess the suggested framework, we used the conventional performance metrics, namely accuracy, recall, precision, and F1-score, which was obtained using the formulae below. Furthermore, a classification problem that determines if a news story is fake or real is tackled similarly to the fake news detection problem. The performance of the fake news detection is computed using the confusion matrix. Here is a quick explanation of these metrics:

$$Accuracy = \frac{TP + TN}{TP + TF + FP + FN}$$

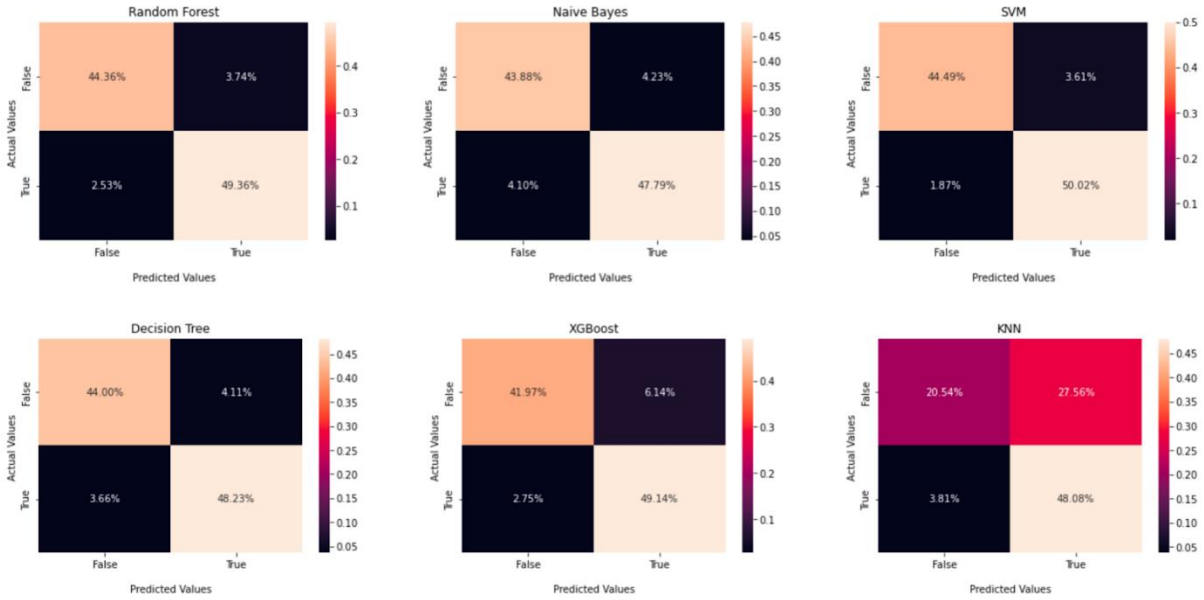
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

where, True positive (TP) = Fake news predicted as fake; True negative (TN) = Real news forecasted as real; False positive (FP) = Fake news predicted as real; False negative (FN) = Real news predicted as fake.

5.3 Confusion Matrix

The confusion matrix was used to represent the outcomes of the dataset analysis employing the six different models. The confusion matrix gives us information about the model's accuracies and also tells about the true positive and false negative values.



5.4 RESULTS

Detecting false news is challenging since news items are always changing; nonetheless, the essential concepts and principles of fake news are highlighted in social media. By examining it in two stages characterization and disclosure the research in this paper focuses on identifying fake news. Kaggle was used to get the dataset. To obtain the highest level of accuracy from the model, we applied the characteristics of tokenization, lemmatization, stop words removal, and NLP in our dataset. The techniques for supervised learning are examined for the detection of bogus news during the discovery phase. Models based on speech features and predictions are used to reveal bogus news. Algorithms with a combined ML accuracy of 94.5% were employed. Several models, including Naive Bayes, Neural Networks, and Support Vector Machines (SVM), were utilized for the project. Out of all of them, we discovered that SVM was the best model, achieving an accuracy of 94.5%.

Lastly, the model classifies the news articles into two classes: real and fake.

6 CONCLUSION

In this study, we created an end-to-end methodology for identifying false news at an early stage by examining both the news article's textual content and the source data. Different supervised learning algorithms are used in the research step to assess the existing techniques for false news detection. Our models' accuracy varies from 65% to 94.5%.

7 RELATED WORK

Fake News Detection Using Machine Learning Approaches.
Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and Machine Learning.
CB-Fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT.
An ensemble machine learning approach through effective feature extraction to classify fake news.

8 REFERENCES

[1] Data cleaning:
https://iaviral.medium.com/what-is-data-cleaning-2ac726312ef4
[2] NLTK:
https://www.nltk.org/
[3] Natural Language Processing (NLP):
https://becominghuman.ai/what-is-natural-language-processing-nlp-16cc0f94858f
[4] Tokenization:
https://www.techtarget.com/searchsecurity/definition/tokenization
[5] Stopword removal:
https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
[6] Lemmatization:
https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
[7] Classification models:
https://iaviral.medium.com/classification-models-cb4ba55c6f4d
[8] Supervised learning:
https://iaviral.medium.com/what-is-supervised-learning-b385ce2b7dfb
[9] Random forest:
https://iaviral.medium.com/what-are-random-forests-an-introduction-to-machine-learning-algorithms-3ea0a9f94f05
[10] SVM:
https://iaviral.medium.com/what-is-support-vector-machine-svm-c1b759db65ba
[11] Decision tree:
https://iaviral.medium.com/what-is-decision-tree-an-introduction-to-machine-learning-algorithms-8d8149d611a6
[12] Naive Bayes:

https://becominghuman.ai/what-is-naive-bayes-1276a384efc
[13] Bayes theorem:
https://en.wikipedia.org/wiki/Bayes%27_theorem
[14] KNN:
https://iaviral.medium.com/what-is-k-nearest-neighbor-knn-811ef165a223
[15] XGBoost:
https://xgboost.readthedocs.io/en/stable/
[16] Gradient boosting algorithm:
https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/
[17] Data visualization:
https://iaviral.medium.com/what-is-data-visualization-7c5bf1b5867c
[18] Line plot:
https://seaborn.pydata.org/generated/seaborn.lineplot.html
[19] Evaluation metrics;
https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226
[20] Confusion matrix:
https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
[21] Fake News Detection Using Machine Learning Approaches
https://doi.org/10.1088/1757-899X/1099/1/012040
[22] Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and Machine Learning:
https://doi.org/10.1007/s42979-021-00775-6
[23] CB-Fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT:
https://doi.org/10.1007/s11042-021-11782-3
[24] An ensemble machine learning approach through effective feature extraction to classify fake news:
https://www.sciencedirect.com/journal/future-generation-computer-systems

[25] Bhatt, Gaurav, et al. "Combining Neural, Statistical and External Features for Fake News Stance Identification." Companion of the Web Conference 2018 on the Web Conference 2018. International World Wide Web Conferences Steering Committee, 2018.
[26] Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2018. Combining neural, statistical and external features for fake news stance identification. Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18.
[27] Rohit Kumar Kaliyar, "Fake News Detection Using a Deep Neural Network", 2018 4th International Conference on Computing Communication and Automation (ICCCA)
[28] Tejaswini Yesugade, Shrikant Kokate, Sarjana Patil, Ritik Varma, Sejal Pawar," Fake News Detection using LSTM", International Research Journal of Engineering and Technology (IRJET), Volume: 08 Issue: 04 | Apr 2021.
[29] Pritika Bahad, Preeti Saxena, Raj Kamal, "Fake News Detection using Bi-directional LSTM-Recurrent Neural Network", INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, ICRTAC 2019
[30] Syed Ishfaq Manzoor, Dr Jimmy Singla, Nikita, "Fake News Detection Using Machine Learning approaches: A systematic Review", Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8
[31] Supanya, Prabhas, "Detecting Fake News with Machine Learning Method",2015 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunication and Information Technology