Accurate Fake News Prediction by Comparing Performance of Machine learning algorithms

Akilandasowmya.G¹, Gauthami Ghadiyaram², M. Pavetha¹, S.P. Hemamalini¹
¹Assistant Professor, Computer and Communication Engineering, Sri Sairam Institute of Technology Chennai, India.
²,³,⁴UG-Computer and Communication Engineering, Sri Sairam Institute of Technology, Chennai, India.

Emails:  akilandasowmya.cce@sairamit.edu.in¹, sit21co041@sairamtap.edu.in², sit21co064@sairamtap.edu.in³, sit21co05@sairamtap.edu.in⁴

Abstract
With the advancement of technology and the widespread use of media there has been an increase, in the circulation of fake news. Unfortunately, some individuals intentionally spread information to manipulate opinion and drive traffic to specific websites. One such instance occurred during the Covid 19 pandemic when misleading rumors started circulating falsely claiming that Covid vaccines were linked to heart attacks and infertility. These baseless claims created hesitancy among people regarding vaccination. To assist individuals in identifying news accurately this paper compares the performance of various machine learning algorithms such as Passive Aggressive Classifier, Decision Tree, Random Forest, Logistic Regression and Naïve Bayes. After evaluating their results, it was determined that the Passive Aggressive Classifier achieved an accuracy rate of 98.2% followed by Naïve Bayes with 96.59% accuracy Random Forest with 96.95% accuracy, Decision Tree with 96.23% accuracy and Logistic Regression with 97.22% accuracy. Based on these findings it can be concluded that the Passive Aggressive Classifier is the algorithm for predicting fake news among all five models tested in this study. The data used for building these machine learning models was obtained from Kaggle website. The primary objective of this research paper is to provide guidance to individuals seeking to choose an algorithm that offers accuracy, in detecting news.

Keywords: Decision Tree; Logistic Regression; Naïve Bayes; Passive Aggressive Classifier; Random Forest;

1. Introduction
The push of social media to the top as a major medium for capturing information has had a huge impact in spreading fake news. The ease with which data can be shared on social media without proper verification has increased the spread of fake news. As a result, it has become increasingly difficult for users to distinguish authentic data from fake news. Additionally, the algorithms used by social media systems tend to engage first with virality over accuracy, promoting sensational or fraudulent content. This has created an environment where fake news can use pull faster and reach more people targeted side, apart from the persistence of misinformation and public acceptance of the facts as to the truth of the equitable property as the ultimate outcome it is important that those who social media uses and policies take proactive measures to prevent the spread of false information. This has blurred the tension between real and fake news, making it harder for consumers to tell what is genuine and what is not because of this as a way to protect you this we use the tool to learn certain rule mechanisms to identify real and gaming issues. The use of gadget learning algorithms to obtain real and fake information is a really important way to prevent misrepresentation. The capabilities of algorithm information, such as deliverability, language style, content consistency and ability to detect fake information can be investigated by studying different methods and comparing analysis with property information a of reliable acceptance.

2. Literature Survey
Dr.S.Gowri et al [1] says about the computational model for detecting fake news using machine learning techniques. This model utilizes a machine learning technique and TF - IDF vectorizer on a
proposed dataset to achieve better efficiency in detecting fake news. This paper includes models based on content, user reaction and source quality. The authors highlight the social impact of fake news and need for improvised method in detecting fake news. Shwetha D.Mahajan [2] discusses the news classification using machine learning techniques, focusing on a news related dataset with various categories such as entertainment, education, sports and politics. They use classifying algorithms and word vectorizing techniques to improve performance of classification model. They compare the model with Naïve Bayes and TF-IDF vectorizer. Results evaluated based on precision, recall, F1 score and accuracy. Jithin Joseph et al [3] discusses a hybrid approach for fake news detection using multinomial voting algorithm. It achieves a precision score of 94% on phony dataset. Algorithms used are Naïve Bayes, Decision tree, Random Forest, K Nearest Neighbours and Support Vector Machine. The impact of fake news on individuals and organizations is highlighted and solved the problem using machine learning techniques. TF-IDF method also used for information retrieval. Dr.S.Madhavi et al [4] proposes fake news detection using machine learning algorithms like Support Vector Machine and Naïve Bayes. Decision tree classifier used in data mining and information analysis. Sentiment analysis and text mining techniques are employed to identify fake news and spams in online platforms. Decision tree and along with property-oriented induction are used for efficient classification. Tao Jiang et al [5] focuses on social media as a major source of news and challenge of identifying fake news. Five machine learning algorithms performance are evaluated (i.e.) Logistic regression, Naïve bayes, Decision tree, Random Forest and Support Vector Machine and used two deep learning algorithms such as CNN and LSTM based on ISOT dataset and KD nugget dataset. They use accuracy, precision, recall, F1 score as evaluation metrics and corrected version of McNemar’s test to determine the model’s performance. Ahm Al Ayub Ahmed et al [6] proposes a systematic literature that focusses on use of machine learning techniques for detecting fake news. Supervised machine learning algorithms are also used. Naïve bayes, Logistic regression, Recurrent neural network, Random Forest and K Nearest Neighbours. It shows how difficult to use supervised learning for detection of fake news and advises to use unsupervised algorithms. Rohit Kumar Kaliyar et al [7] proposes a Bert based deep learning approach for fake news detection in social media. They combine the single layer of deep convolution neural network with Bert to handle ambiguity and improve classification. Experiments are performed on CNN and LSTM algorithms along with pretrained word embedded techniques such as BERT and GloVe. The performances are analysed and compared with benchmark results. Hagar Saleh et al [8] addresses the issue of fake news and its impact on social cohesiveness and political polarization. This paper aims to provide optimized convolution neural network model (OPCNN-Fake) for detection of fake news and compares the results with RNN and LSTM. NGram, TF-IDF and Glove word embedding are used to extract features from datasets. Grid search and hyper opt optimization techniques used to optimize parameters of ML and DL models. Faraz Ahmad et al [9] proposes deep learning methods to detect fake news and uses different types of neural networks such as CNN, LSTM and feed forward and were trained and evaluated on labelled dataset of real and fake news. The models performed effectively in detecting fake news with convolutional and LSTM showing the most effectiveness. Supanya Aphiwongsophon et al [10] says machine learning techniques as effective means to predict fake news. Naïve bayes, Neural network and Support vector machine have been used to predict fake news. Normalization method is used for cleaning the data. The results show that Naïve Bayes shows that accuracy 96.08% and Neural network and Support machine shows the accuracy of 99.90%. The results provide the highest accuracy for fake news detection. Sheng How Kong et al [11] aims to apply Natural language processing methods and deep learning methods to detect fake news. Tensor flow with built in Kera’s is used as a framework. Model evaluation is done using recall and precision metrics. This study aims to compare the performance of different deep learning algorithms such as RNN and LSTM. Sherry Girgis et al [12] proposes RNN models such as vanilla, GRU, LSTM to detect fake news. They were evaluated on LAIR dataset and GRU provided the
highest accuracy .217 followed by LSTM having 0.216 and Vanilla having 0.215. To increase the accuracy a hybrid model by combining GRU and CNN is applied to the dataset. Chaitra.K.Hiramath et al [13] proposes to detect fake news by using Deep neural network, Logistic regression ,Naïve bayes ,Support vector machine and Random Forest. The framework is based on java system and uses Netbeans and SQL for database. The performance comparison shows DNN has higher accuracy than others. The dataset used to train the models are PHEME dataset and Liar dataset. Ethar Qawasmeh et al [14] proposes to challenge identifying the fake news in online communication platforms. The author proposes the automatic identification model based on bidirectional LSTM. The dataset used is FNC-1 dataset derived from Craig Silverman. The proposed model achieves a accuracy of 85.3%. Junaed Younus Khan [15] focusses on fake news in social media and its potential impacts. Various pretrained language models, deep learning models and traditional models for fake news detection were used. BERT was found to provide the best results. Models were all based on article length and article topics. They used three different types of datasets the largest and most diversified ones.

3. Algorithms Used in This Paper

3.1 Logistic Regression

Logistic regression is one of the immensely useful statistical techniques in use for evaluating information. Using mathematical principles, it can establish relationships between variables. It is often used in classification and predictive analytics because it correctly predicts the probability of an event, yes or no, on the basis of a given set of independent variables. It is, in fact, an extension of the linear regression techniques. Logistic regression function

\[ f(x) = \frac{1}{1 + e^{-x}} \]

3.2 Naïve Bayes

One of the most frequent algorithms applied in text classification is probably Naïve Bayes. It's based on the famous Bayes theorem, one of the ideas that comprise the core of probabilistic machine learning. The unique property of this algorithm is that it learns to distinguish between classes without learning the most significant features, unlike discriminative classifiers, like logistic regression.

\[ P(Y / X) = P(X and Y) / P(X) \]

3.3 Random Forest

Random forests are powerful machine learning algorithms that consist of many decision trees. Their main function is to classify email and news with a good degree of accuracy as either spam or not spam. This algorithm will turn out with very high precision and has significant feature importance, hence making the cross-validation or a separate test set in the estimation of test errors null.

\[ m = \text{sqrt root}(p) \]

3.4 Decision Tree

The Decision Tree has a strong classification and regression ability, both of which help in supervised learning. But it often deals with a problem of course. In this technique, the data can be branched efficiently under some conditions of features, having a tree-like structure. There exist two faculties on the Decision Tree: the Decision Node faculty and the Leaf Node faculty. While the former would be the decision maker, the latter would be the result obtained from that decision and hence would not reach any further funding. The building of the tree was done using the CART algorithm, which stands for `Classification and Regression Tree algorithm'.

\[ G(t) = 1 - \sum_j = 1 C(p_i)2 \] where C is the number of classes, and p is the proportion of class i instances in node t.

3.5 Passive Aggressive Classifier

The passive aggressive algorithm is a relevant tool of machine learning, heavily utilized in the online field, principally for classification problems. It turns out to be very useful with large amounts of data, often happening in big data applications. In these instances, the size of the data might be small enough that it is not possible to train the whole dataset. Therefore, the passive aggressive algorithm will become one of the prime options. In summary, the passive aggressive classifier is one of the great algorithms for systems dealing with a stream of data.

\[ w' = w + \Delta w \]
\[ b' = b + \Delta b \]

where w ’ and b ’ are the updated weight vector and bias term

4. Dataset
Kaggle is an international information technology and machine learning competition platform that gives access to a wide variety of datasets for diverse programs. Right here, statistics scientists and system getting to know practitioners can locate and put-up datasets, as well as discover and construct fashions in an internet-based statistics technological know-how environment. They also can collaborate with different experts inside the subject and participate in competitions to clear up facts technology challenges.

By way of utilizing the energy of crowdsourcing and the understanding of statistics scientists, Kaggle enables individuals and agencies to address complex troubles and drive innovation inside the field of statistics technological know-how and system studying. In this venture we’ve use fake news shown in Figure 1 and real news shown in Figure 2 dataset from Kaggle, shown in Table 1.

![Figure 1 Fake News](image1)

![Figure 2 Real News](image2)
Table 1 Comparison of Five Algorithms

<table>
<thead>
<tr>
<th></th>
<th>Logistic Regression</th>
<th>Naïve Bayes</th>
<th>Random Forest</th>
<th>Decision Tree</th>
<th>Passive Aggressive Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>Logistic regression is a statistical method which is used in analysis technique which is used in analysis technique</td>
<td>The way that naïve bayes functions is by utilizing the principle of conditional probability.</td>
<td>The random forest technique utilizes the power of multiple decision trees to produce one unified outcome.</td>
<td>The decision tree serves the dual purposes of classification and regression.</td>
<td>The Passive Aggressive algorithm boasts the highest accuracy among all others, making it</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td>Linear</td>
<td>Probabilistic</td>
<td>Ensemble</td>
<td>Tree-based</td>
<td>Online learning</td>
</tr>
<tr>
<td><strong>Interpretability</strong></td>
<td>Moderate</td>
<td>High</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Complexity</strong></td>
<td>Low moderate</td>
<td>Low</td>
<td>High moderate</td>
<td>Moderate</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Training Speed</strong></td>
<td>High</td>
<td>High</td>
<td>High moderate</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td><strong>Feature Importance</strong></td>
<td>Varies</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Varies</td>
</tr>
<tr>
<td><strong>Memory Usage</strong></td>
<td>Low</td>
<td>Low</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Low</td>
</tr>
</tbody>
</table>

5. Proposed Method

![Figure 3 Block Diagram](https://irjaeh.com/doi/10.47392/IRJAEH.2024.0299)

**Figure 3** Block Diagram
We have imported both fake and real dataset from Kaggle website. Imported the necessary libraries such as NumPy, Pandas, Matplotlib, re. Chose the algorithms Naïve Bayes, Logistic Regression, Passive aggressive classifier, Random Forest, Decision tree. The next step is the data pre-processing, where data is cleaned to remove the punctuation marks, html contents and stop words. Next step in pre-processing is Lemmatization where words are converted to root word. N gram analysis is done to test the frequency of words. Tokenization is done to decrease the word limit. As a greater number of words can decrease the execution speed of the process. Split the data into training and testing data to find the accuracy of the model. After data pre-processing train the data with the training data. Test the model with the testing data to predict the accuracy of the model. At last compare the results of five algorithms and select the best one, shown in Figure 3.

6. Implementation

Bar chart Results after comparing the Five Algorithms as shown in Table 2.

Table 2 Bar Chart Results After Comparing the Five Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>96.59</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>97.22</td>
</tr>
<tr>
<td>Decision tree</td>
<td>96.23</td>
</tr>
<tr>
<td>Random Forest</td>
<td>96.59</td>
</tr>
<tr>
<td>Passive aggressive classifier</td>
<td>98.02</td>
</tr>
</tbody>
</table>

Conclusion

We have detected the fake news from fake and real dataset from Kaggle by using five machine learning algorithms and compared their results. By the end of the implementation, we found Passive aggressive classifier to provide the highest accuracy. Passive aggressive classifier provided accuracy of 98.02%.

Logistic regression provided accuracy of 97.22%, Decision tree provided accuracy of 96.23%, Random Forest Naïve bayes provided accuracy of 96.59%. Passive aggressive classifier turned out to be the best algorithm.

Future Work

We have done these predictions on a predefined dataset so we are trying to extend our predictions in real time data like the news websites. We are even trying to predict fake audio, video, images and integrating them into an app /web which are easily accessible by people and preventing them blindly trusting the unknown messages.

References


[6]. Alim Al Ayub Ahmed Ayman Aljarbouh
Praveen Kumar Donepudi Myung Suh Choi, Detecting Fake News using Machine Learning: A Systematic Literature Review, UST global [7].


