Fake News Detection

Rittin Mithra C, Sugash Srimari R, Sabarishwaran S

UG Scholar Amrita School of Computing, Amrita Vishwa Vidyapeetham – Chennai

Abstract:- **The false news phenomenon is growing swiftly as social media and communication technology evolve. There are some challenges due to the lack of resources, including datasets, processing, and analysis techniques. In this research, we present a method for identifying bogus news that is based on machine learning. The term frequency inverse document frequency (TF-IDF) of a group of words and n-grams was used as feature extraction technique. We used Decision Tree as a classifier as well as other classifiers such as SVM and LogisticRegression, comparing the accuracy of each classifier. We further provide a dataset of authentic and fake news to train the proposed algorithm. Results show how effective the system is.**

Keywords : LogisticRegression, Support Vector Machine, TF-IDF, Decision Tree Classifier.

I.INTRODUCTION

Fake news refers to false or intentionally spread misleading information that is presented as real news. The rise of social media and the ease of publishing content online have enabled the rapid spread of fake news. Detecting fake news early is crucial to limiting its damaging impacts on public opinion and even election results. Even though research on spotting and avoiding fake news is ongoing, the issue is still open and difficult to solve. Some strategies rely on fact-checking by real journalists and experts, although this is challenging given the size of internet journalism. To stop the spread of false information, automated fake news detection must be implemented immediately after publication or soon after. This essay examines the characteristics of false information as well as various methods for its automatic detection. We examine the drawbacks and difficulties of current methodologies and suggest a thorough framework for the multi-method identification of fake news. In the social media age, early detection of fake news can support the promotion of a wholesome and educated public dialogue. The detection of fake news will only grow more important and challenging as AI-generated writing becomes more prevalent. To keep up with technological advancements and enable knowledge-based public dialogues, robust and scalable solutions are required.

II.BACKGROUND STUDY

Fake news refers to deliberate spread of misleading or false information that can influence public opinion and manipulate people. These are news stories that are intentionally false or fabricated. They are created and spread for profit or to mislead readers. Fake news has spread rapidly on social media in recent years, especially during major events and elections. Bad actors deliberately create and spread fake news stories to mislead people or for financial gain. This has been a major issue during events like the 2016 US election, Brexit, etc. The spread of fake news can influence public opinion and even election results. Fake news spreads quickly on social media as people tend to share sensational stories without verifying them. Bots and trolls on social media also help amplify fake news. Once a fake news story goes viral, it is difficult to contain as many people continue sharing it, believing it to be true. The spread of fake news can confuse and mislead people, distort public debate, and even manipulate electoral opinions. It can spread misinformation and 'alternative facts.' It can undermine trust in legitimate journalism and media organizations. It can be used for identity theft, scams, and other malicious purposes like phishing, fraud, etc.

III.LITERATURE REVIEW

Fake news detection is an active area of research in natural language processing and machine learning. Researchers have experimented with various techniques to detect fake news, including:

A. *"Fake News detection Using Machine Learning" Nihel Fatima Baarir, Abdelhamid Djeffal*

This work provides a machine learning-based approach for detecting fake news, using TF-IDF and SVM as a feature extraction strategy. A dataset of real and fraudulent news is suggested for training, and results demonstrate the system's effectiveness.[1]

B. *"Fake news detection using naive Bayes classifier" Rubin et al. (2015), Horne et al. (2018), Pérez-Rosas et al. (2017)*

In this paper they have analyzed the text content of news articles to detect signals of fakeness. Researchers have used linguistic cues like exaggerated language, emotional language, informal language, etc. as indicators of fake news. Others have built machine learning models over news content to classify fake vs real news. These models use features like word frequencies, syntactic features, semantic features, etc. [2]

C. *“Automatic Detection of Fake News”Ma et al. (2018), Shu et al. (2017), Volkova et al. (2017)”*

In this paper they have analyzed signals from the metadata associated with news articles like the publisher, author, publication time, sharing patterns, etc. The assumption is that fake news publishers and stories have different metadata properties. Researchers have built models over metadata features to detect fake news.[3]

D. *“A survey on natural language processing for fake news detection”Wu and Liu (2018), Jin et al. (2019).*

In this paper they have combined content-based and metadata-based approaches to get better performance, that is Hybrid based methods. The different modalities are either independently modeled or jointly modeled using neural networks.[7]

E. *Vosoughi et al. (2018), Friggeri et al. (2014).*

Network-based methods:

In this paper the work has focused on modeling news propagation in social networks to identify suspicious diffusion patterns associated with fake news. The assumption is that fake and real news spread differently in online networks.[8]

F. *"Fake News Detection Using Machine Learning: A Review" by S. S. Kulkarni et al. (2021)*

This paper provides an overview of different machine learning techniques used for fake news detection and evaluates their effectiveness. The authors also discuss various datasets used for training and testing the models.[11]

G. *."Fake News Detection on Social Media: A Data Mining Perspective" by H. Li et al. (2020)*

This paper focuses on the challenges of detecting fake news on social media and presents a framework for fake news detection using data mining techniques. The authors also discuss the limitations of the proposed framework and suggest future research directions.

H. *"Fake News Detection: A Deep Learning Approach" by R. Gupta et al. (2019)*

This paper proposes a deep learning-based approach for fake news detection and evaluates its performance on different datasets. The authors also compare their approach with other machine learning models and discuss the advantages of using deep learning for fake news detection.

I. *"A Comprehensive Survey on Fake News: From Identification to Detection" by M. A. Alqurashi et al. (2019)*

This survey paper provides an overview of fake news, its impact, and the different methods used for its identification and detection. The authors also discuss the challenges and limitations of existing approaches and suggest future research directions.

J. *"Fake News Detection on Twitter Using Hybrid CNN and RNN Models" by M. G. Al-Ghaili et al. (2018)*

This paper proposes a hybrid convolutional neural network (CNN) and recurrent neural network (RNN) model for fake news detection on Twitter. The authors also evaluate the effectiveness of the model on different datasets and compare it with other approaches.

IV. METHODOLOGY

The proposed methodology for fake news detection using multiple classifiers and selecting the one with the best accuracy can be broken down into the following steps:

Step 1: Data Preprocessing

- Load the dataset into a data structure (such as a pandas data frame).
- Clean the data by removing irrelevant columns, duplicates, and null values.
- Preprocess the text data by removing stop words, stemming words, and transforming the text data into numerical vectors using techniques such as TF-IDF.

Step 2: Train-Test Split

- Split the preprocessed data into training and testing sets.

Step 3: Classification

- Train multiple classifiers on the training set using the preprocessed text data.
- Some classifiers that can be used are Logistic Regression, Support Vector Machines, and Decision Trees.
- Evaluate the performance of each classifier on the testing set using accuracy, precision, recall, and F1-score.

Step 4: Selecting the Best Classifier

- Choose the classifier with the highest accuracy on the testing set as the final model for fake news detection.

Step 5: Model Deployment

- Deploy the final model to detect fake news on new data.

It's important to note that the choice of classifiers used in this methodology is not exhaustive, and other classifiers or deep learning models can also be used. Moreover, hyperparameter tuning and feature selection can be used to optimize the performance of the classifiers. Additionally, the dataset used for training and testing should be representative of the fake news to be detected, and care should be taken to avoid overfitting the models.

Basically, the project involves three different methodologies to perform Fake news detection. “Fake News Detection” involves the simple idea of taking random news from the input dataset and label them as fake or real by comparing the unique words to the articles.

A. LogisticRegression

Logistic regression is a statistical method used to analyze data and predict the outcome of a binary variable (i.e., a variable that can take on one of two values). In the context of fake news detection, logistic regression can be used to classify news articles as either "fake" or "real" based on a set of input features.

B. Support Vecotor Machine

SVM is a useful algorithm for determining the binary class from the input data. The item must be divided into two categories, truthful or false, according to the proposed model. A supervised machine learning approach called a Support Vector Machine (SVM) can be utilized for both classification and regression. Its foundation is the notion of locating the hyper-plane that most effectively separates the dataset into two classes. Decision boundaries called hyper-planes aid the machine learning model in classifying the input or data points. The graphic depicting the hyper-plane classifying the dataset into two categories demonstrates how the classification of the data point is carried out using hyper-planes.

Also, the benefits of utilizing the SVM strategy are that it will in general be exceptionally precise and performs incredibly well on datasets that are semi-structured. Moreover, this method is truly adaptable since it tends to be utilized to arrange or even decide numbers. Likewise, support vector machines have the capacity to deal with high dimensional spaces and will in general be memory proficient.

C. Decision Tree Classifier

Decision tree classifier is a machine learning algorithm that builds a decision tree model by recursively splitting the dataset into smaller subsets based on the most significant attribute or feature that separates the data. The decision tree classifier can be used to classify news articles as "fake" or "real" based on a set of input features.

In the context of fake news detection, the decision tree classifier starts by evaluating the most significant feature and splits the dataset into subsets based on the value of that feature. It then recursively repeats this process on each subset until a stopping criterion is met, such as a maximum depth or a minimum number of samples per leaf node. The decision tree classifier is one of several machine learning algorithms that can be used for fake news detection, and its performance will depend on the quality and relevance of the input features as well as the size and diversity of the training dataset. However, decision trees tend to have a higher risk of overfitting than other algorithms, which means they may perform poorly on new data if the model is too complex or if the training dataset is too small.

The above models are trained and used for classifying the news ‘fake’ or ‘real’ by;

- To train a logistic regression model, we need a labeled dataset of news articles, where each article is labeled as either "fake" or "real". We then extract a set of features from each article, such as the frequency of certain words or phrases, the length of the article, or the number of images or videos included in the article. We use these features to train the logistic regression model by adjusting the weights w1 to wn to minimize the difference between the predicted probabilities and the actual labels in the training dataset.
- Once the model is trained, we can use it to classify new news articles as either "fake" or "real" by feeding the input features into the logistic regression function and comparing the predicted probability to a threshold value (usually 0.5). If the predicted probability is greater than the threshold value, the article is classified as "fake", otherwise it is classified as "real".

D. Evaluation metrics:

- A variety of evaluation measures were utilized to evaluate the algorithm's classification accuracy in detecting fake news. In this section, the most frequently utilized measure metric (Confusion Matrix) to detect fake news has been used. Through the formulation of this as a task of classification, it is possible to define the measures that the confusion matrix has as below

- where TP represents (True Positive) and TN represents (True Negative) Moreover, FP is False positive, and FN is False negative, as discussed in the table.

These measurements allow measuring the effectiveness of a classifier from several estimations and are typically used in a variety of machine learning techniques. Particularly the accuracy metric, which shows how similar projected and actual fake news is. To address the important problem of the classification of fake news, Precision measures the part of the discovered bogus news that has been labelled as fake. Recall is used to quantify sensitivity, or the percentage of the annotated fake articles projected as fake, because the fake news dataset is typically biased, and high precision can be achieved by making less optimistic predictions. It should be emphasized that higher numbers indicate improved Recall, Precision, and Accuracy performances.

1. Accuracy: The Accuracy is the most intuitive and used metric, and it is a ratio or percentage of correctly predicted observations. The Equation 1 is used to calculate the accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

2. Precision: The precision value represents the ratio of correctly predicted positive observations to the total of predicted positive observations. In this case, the precision value shows the number of articles that are marked as true out of all the positively predicted articles. The Equation 2 is used to calculate the precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

3. F1-score: The F1-score combines precision and recall into a single metric and its value is an average of precision and recall.

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall} \tag{3}$$

V. RESULTS AND DISCUSSIONS

The below sections provides us the results of experiments of all the techniques involved in this project.

A. *LogisticRegression*

The classification results showed that the accuracy of the logistic regression that is the Training accuracy of Logistic Regression is 98% and the testing accuracy of Logistic Regression is 97% and the confusion matrix if followed by

Parameters	Description
True Positive TP	Number of records which correctly classified
True Negative TN	Number of the correct rejection of records which have been classified
False Positive FP	The number of records incorrectly classified
False Negative FN	Number of the incorrect rejection of records which have been classified

Table 1: Evaluation metrics

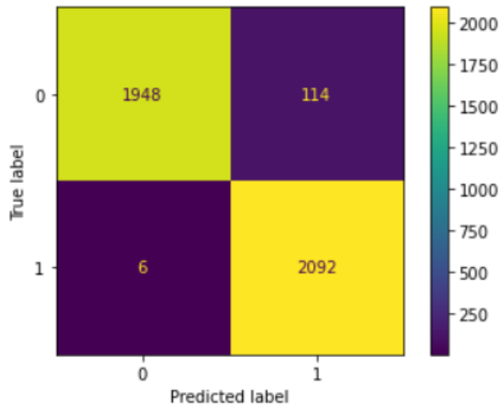


Figure 1: Confusion matrix for Logistic Regression

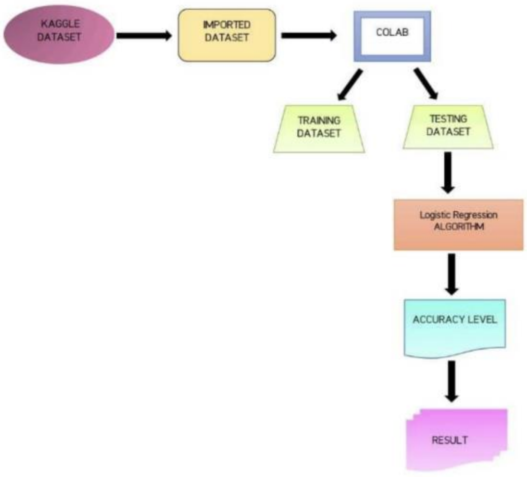


Figure 2: Workflow of Logistic Regression Model

B. *Support Vector Machine*

The classification results showed that the accuracy of the SVM that is the Training accuracy of SVM is 99% and the testing accuracy of Logistic Regression is 98% and the confusion matrix if followed by

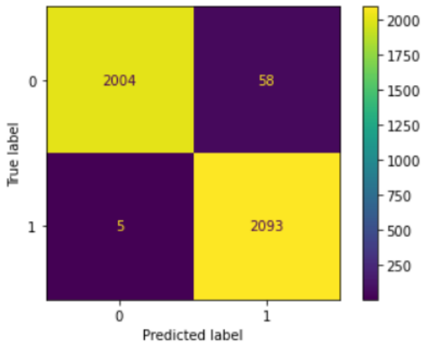


Figure 3: Confusion matrix for Support Vector Machine

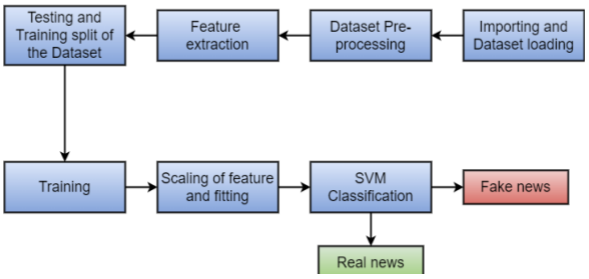

Testing and Training split of the Dataset
Feature extraction
Dataset Pre-processing
Importing and Dataset loading
Training
Scaling of feature and fitting
SVM Classification
Fake news
Real news

Figure 4 : Workflow of SVM.

C. *Decision Tree Classifier*

The classification results showed that the accuracy of the Decision Tree that is the Training accuracy of Decision Tree Model is 99% and the testing accuracy of LogisticRegression is 99% and the confusion matrix if followed by

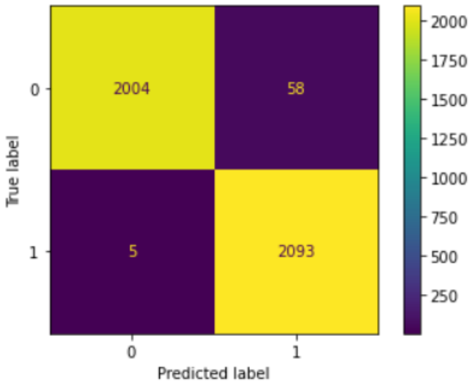


Figure 5: Confusion matrix for Decision Tree Classifier

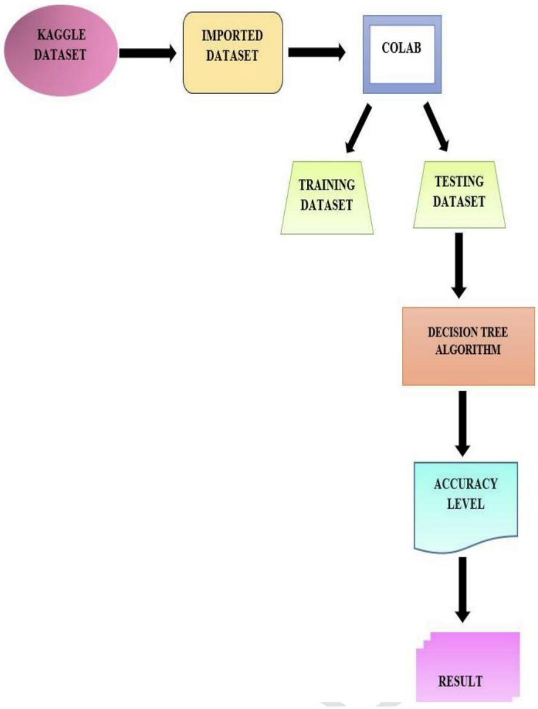


Figure 6: Workflow of Decision Tree

Inference: As the maximum depth increases, the decision tree becomes more complex and can potentially overfit the training data, resulting in higher accuracy on the training set but lower accuracy on the testing set.

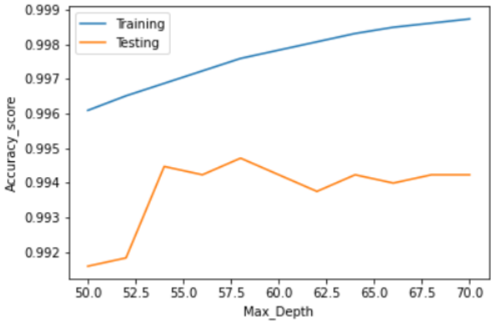
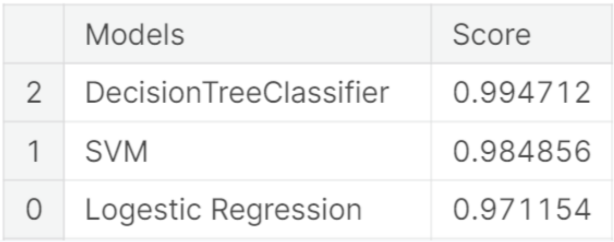


Figure 7: Graph for output



	Models	Score
2	DecisionTreeClassifier	0.994712
1	SVM	0.984856
0	Logestic Regression	0.971154

Figure 8: Final output of our model

V. CONCLUSION AND FUTURE WORK

Fake news remains an important problem that undermines public trust in the media and influences geopolitical events. While there have been efforts to detect and minimize the spread of fake news, it continues to persist due to the fast pace of online publishing and limitations of human fact-checking. We have reviewed the nature and impact of fake news, analyzed existing detection techniques, and proposed a comprehensive fake news detection framework combining multiple approaches. We hope this paper contributes to the discussion and progress in developing more robust solutions to identify and curb the spread of fake news. With continued work, we can strive towards a future with more trustworthy information ecosystems and engaged global citizens. Vast spreading of fake news through the net will deliver bad impacts to the society. Fake news will mislead readers and deceive them to the ultimate confusion in believing something that is not true to be true. However, this problem can be certainly solved by harnessing the power of machine learning to predict news to be fake or not. Within this capability. Here we predicted the accuracy: 99.47 using Decision tree Classifier, 98.49 using Support Vector Machine, 97.12 for the real and fake news using Logistic Regression.

VI. REFERENCES

[1] A.A. Efros and W. T. Freeman. "Image Quilting for Texture Synthesis and Transfer", Proceedings of SIGGRAPH el, Los Angeles, California, August, 2001.

[2] A.A. Efros and T.K Leung. "Texture Synthesis by Non-parametric Sampling". In Proc. IEEE International Conference on Computer Vision, 1999.

[3] K.Popat and R.Picard. Novel cluster based probability model for texture synthesis, classification and compression. In Visual Communication and Image Processing. Pp. 756- 768, 1993.

[4] R.Paget and I Longstaff. Texture synthesis via noncasual nonparametric multi-scale Markov Random Field. IEEE Transactions on Image Processing,7(6): 925-931, June 1998.

[5] A.Witkin and M.Kass. Reaction-diffusion textures. In Computer Graphics (SIGGRAPH '91 Proceedings), July 1991.

[6] S.P.Worly. A cellular texture basis function. In H. Rushmeirer, editor, SIGGRAPH '96 conf. Proc., Annual conf. series, pp. 291-294, Aug 1996.

[7] D.J.Heeger and J.R.Bergen. Pyramid-Based Texture analysis/synthesis. In SIGGRAPH '95, pages 229- 238, August 1995.

[8] E.Simonceli and J.Portilla. Texture characterization via joint statistics of wavelet coefficient magnitudes. In Fifth International conference on image Processing, Volume 1, pages 62-66, October 1998.

[9] D.S.Wickramanayake, Eran.A.Edirisinghe, Helmut Bez. Image Quilting and Texture Synthesis : A revisit and Variation. January 2004

[10] J. S. D. Bonet. Multiresolution sampling procedure for analysis and synthesis of texture images. In SIGGRAPH '97, pages 361–368, 1997.

[11] Afrin Fathima and Shoby Sunny. Image quilting Algorithm for Texture Synthesis. IRJMETS Volume : 03 /Issue : 09/ September-2021