ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Impact Factor: 8.206

Volume 8, Issue 8, August 2025

© 2025 IJMRSET | Volume 8, Issue 8, August 2025| DOI:10.15680/IJMRSET.2025.0808131

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

ENHANCING THE DETECTION OF FAKE NEWS IN SOCIAL MEDIA BASED ON MACHINE LEARNING MODELS

Barnali Chakraborty, Sahana Naik

Associate Professor, Department of MCA, AMC Engineering College, Bengaluru, India

Student, Department of MCA, AMC Engineering College, Bengaluru, India

ABSTRACT: Inaccurate or misleading reporting, sometimes known as "fake news," has become a significant problem in modern society. Fake news is a severe problem in the digital landscape, where anyone can publish their thoughts as though they were facts. However, research into what social media users take in is still in its infancy, and ongoing efforts aim to learn more about how social media users can identify false information. Discovering a reliable method for identifying fake news has been the biggest obstacle. Using supervised learning techniques that rely on labelled data to determine whether or not a piece of text is genuine helps us become aware of such stories. This study uses multiple methods, such as logistic regression, the XGBoost model, and the Random Forest model, to examine the characteristics of news articles and determine whether they are authentic. These models are typically trained on labelled data, with additional features derived from CNNURL and Reddit data. The goal is to create more reliable sources of information and reduce the spread of misinformation by creating models that can accurately distinguish between real and fake news.

KEYWORDS: Logistic regression, Extreme Gradient boosting, Random forest, Word to Vector, FakeNews detection.

I. INTRODUCTION

The widespread distribution of false information is a significant problem in today's connected world, with negative consequences for individuals, businesses, and communities [1,2]. Combating the spread of false information is an urgent matter [3]. Over the past few years, machine learning algorithms have been used to improve fake news detection. Our goal in this project was to use three well-known machine learning algorithms—logistic regression, XGBoost, and random forest—to enhance the detection of fake news across two platforms: Reddit and CNN.

Many well-known machine learning algorithms can examine textual data and identify patterns indicating whether a news article is authentic. The efficacy of these algorithms in detecting fake news from various sources has been demonstrated across a wide range of studies. The goal of this project is to improve the detection of fake news on Reddit and CNN by using logistic regression, XGBoost, and random forest algorithms to extract features from the news articles, such as text length, source credibility, and language complexity, and then using these features to train and test the models. This research aims to show that these algorithms help spot fake news and to look into their potential for enhancing this accuracy. Furthermore, the project will stress the significance of linguistic complexity and the trustworthiness of sources in identifying fake news.

II. RESEARCH AIM

In order to create influential and trustworthy machine learning models for automatically distinguishing between real and fake news, this study aims to examine fake and real news using labelled data with more features extracted from CNN-URL and Reddit.com. We aim to train models that can accurately identify fake news and stop its spread by analyzing various components of news articles, including linguistic patterns, text structure, and the social media context. We train our models on large datasets with highquality annotations by utilizing labelled data from trustworthy sources like Kaggle and additional features from CNN-URL and Reddit.com. To create a more reliable and well-informed information ecosystem, we must develop machine learning models to categorize news articles accurately. Such models can aid in the tussle against the spread of misinformation, increase public trust in media and political

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

institutions, and facilitate well-informed decisionmaking in various fields by identifying and thwarting the spread of fake news.

III. MACHINE-LEARNING TECHNIQUES

Modeling the relationship between a collection of independent values (inputs) and a set of dependent values is the main goal in many different technology areas (outputs). If such a statistical approach can be created, then it will be possible to infer the required quantities from the ones that may be seen [1,4].

A. Types of Machine Learning

Machine learning may be broken down into three basic categories: supervised, unsupervised, and reinforcement learning [1,5].

In order to learn, supervised methods require training data that includes the value of each input variable's output. In this way, training data consists of pairs of input and output values. The model uses training data to predict output variables based on input variables [5, 6], [7].

Unlike supervised algorithms with a single output value for each input, unsupervised algorithms use a range of values as inputs. The unsupervised algorithm investigates the input data for previously unsuspected connections or patterns [5, 6, 7].

Iterative self-learning methods include reinforcement learning. In this approach, the inputs are the first stages, and the results are the outputs. The model begins training early on, returns with potential solutions, and is punished or rewarded for its performance. After the model generates all possible results, the best one will be selected carefully using the highest reward per solution as a criterion [5, 6, 7].

1). XGBoost

When it comes to solving regression and classification problems, the machine learning algorithm known as extreme Gradient Boosting (XGBoost) is highly regarded for its efficacy. It is a type of ensemble learning in which several decision trees are used together to boost prediction quality. XGBoost iteratively trains a forest of decision trees using the gradient boosting method to achieve the lowest possible prediction error. To prevent overfitting and simplify models, the XGBoost algorithm is an additional improvement to this method. XGBoost excels in scalability, efficiency, and flexibility compared to other machine learning algorithms. It has been used in many settings, from image classification to NLP to recommendation systems, and it can process massive datasets with high-dimensional features.

2). Logistic Regression

Relying on logistic regression to classify linear and nonlinear data is a great idea [8]. Using the Logistic regression machine learning technique, the authors [9] developed a model that can identify false stories. Logistic regression is a standard tool when modelling data with binary responses (1,0), where 1 is typically indicative of success and 0 of failure. This work uses a binary system to label articles as "real" or "fake," with 1 accurately denoting reporting and 0 denoting false news.

3). Random Forest Model

In order to boost prediction accuracy and lower overfitting, the Random Forest model employs an ensemble learning approach known as constructing multiple decision trees from distinct features and data subsets. A vote determines the final forecast among all of the trees [10]. The algorithm can quickly and easily scale and estimates the significance of individual features. Because of their high accuracy, robustness, and interpretability, random forest models have been implemented in many fields, including classification, regression, bioinformatics, and finance.

IV. NATURAL LANGUAGE PROCESSING TECHNIQUES

Natural Language Processing (NLP) can be found in the realm of AI. Because of this, machines can now read, comprehend, and derive meaning from human languages as quickly as a human can (like English, Spanish, Italian, and so on). There are two main branches of natural language processing: NL generation and NL understanding. For

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

analytical purposes, this method can convert raw data into a more organized format [11]. NLP facilitates communication between the fields of data science and language.

The main problem with NLP is that it might make the language more difficult. Mastering the art of language manipulation and comprehension is a mission of paramount importance. This can be avoided by employing various approaches to solving a wide range of problems before attempting to piece together the big picture [11]. In the following sections, we will examine several popular methods employed by NLP for managing the vocabulary of terms.

A. Word2Vec (Word to Vector)

As a tool for uncovering hidden patterns and associations in text, word2vec can be helpful in the fight against fake news. It is possible to determine whether or not a given article or piece of content is likely fake or misleading by analyzing the semantic similarities between words. One strategy involves constructing an educated language model with word2vec using a massive corpus of text containing genuine and fabricated news articles. The model is trained to recognize word patterns and relationships, enabling it to distinguish between credible and fabricated reports. A different strategy involves comparing the article's word embedding to a database of known fake news articles using word2vec to generate the embedding. Comparison of the report's embedding's to those of other, more well-known examples of fake news can help establish whether or not the article is also a hoax. Word2vec can recognize the types of sensational language (our news datasets contain the sensational language), inflated claims, and conspiracy theories typical of fake news articles. Constructing a more precise model for identifying fake news is possible by examining the context in which these words appear. As a whole, word2vec's use in detecting fake news thoroughly examines the language employed in the articles in question. It identifies patterns and relationships between words that may indicate the articles' falsity. Using this method, we can increase the reliability of identifying false news stories and lessen the prevalence of misinformation.

B. Removal of stop word

In natural language processing (NLP), eliminating stop words is a standard method for bettering the precision and productivity of text analysis. Common words like "the," "and," "of," and "is" are examples of "stop words," which are used frequently but do not have much meaning on their own [12]. These words frequently occur in the text but contribute little to the analysis, so they are typically omitted beforehand. Stop words are removed from text to make it easier to read and focus on the essential words and phrases, which improves the results of subsequent NLP tasks like sentiment analysis and topic modelling [14]. NLP applications can perform text analysis with greater precision and efficiency by eliminating these words. To remove all of the stop words, we import "stopword" module from NLTK library and load the list of English stopword. We then split sentences into words, filter out the stopwords using a list comprehension, and join the remaining words into strings.

C. Lemmatization

For natural language processing (NLP), reducing words to their lemma (i.e., their root form) is a common technique. The English words "am", "is", and "are" can all be shortened to the lemma "be". Like "run," "running," and "ran," all three can be distilled into "run" as a lemma [15]. Lemmatization is a method for standardizing text and reducing the number of unique words that need to be processed by reducing them to their base form. As a result, this can boost efficiency in subsequent NLP steps, such as topic modelling or sentiment analysis. Languages with rich morphologies benefit significantly from lemmatization, in which a single word can take on various forms depending on the speaker's meaning. Lemmatization is thus a powerful method for enhancing the precision and efficacy of text analysis in NLP programs. To lemmatize any sentences of news, we create an instance of the WordNetLemmatizer() and call the lemmatize() function on the list of words.

V. RELATED WORK

Rubin et al. [16] propose network analysis and machine learning as a means of determining an article's veracity. This study compares and contrasts several methods for gauging the credibility of claims and offers guidelines for developing a hybrid system. Network and linguistic practices have received the most attention from the two critical approaches [10]. In order to determine whether or not a given piece of text is genuine, the linguistic system employs machine learning techniques. Instead, the network method looks at the queries and metadata. Grouping words and finding n-grams, probabilistic context-free grammars for rule-based classification, and discourse analysis are standard linguistic

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

methodologies. Social media engagement classifies hoaxes. In conclusion, the authors describe a model that blends network and linguistic approaches to categorize linguistic qualities utilizing several layers to improve accuracy [2].

In a recent research [17], fake news characterizations on social belief and data mining, presumption, and assessment metrics were examined in a survey. Incorrect information damages society and tarnishes corporations. It is difficult to tell if a news article is fake just by reading the text. So, it is vital to double-check some extra details. The news articles are separated into two categories for this research: characterization and detection. Technical changes in readers' habits are discussed during the characterization stage. People who read news online rely on social media sites like Facebook and Twitter much more than traditional media like television and newspapers. As a result, social media has become a significant conduit for disseminating potentially misleading news stories. The process is quick, cheap, and simple to grasp. Misinformation is spread on social media for various reasons, including financial gain and political advantage. Fake news may influence users. We employ machine learning methods and feature extraction to identify the article's authenticity. [17].

Hai et al. [18] suggest connecting numerous review "spam detection" tasks with unlabeled data to solve the lack of labeled data. Feature extraction and supervised classification need much-labeled data and human judgment from domain experts. Logistic Regression (LR) offers an innovative solution. This technique automatically obtains information while developing a model through knowledgesharing. This model used a laplacian regulariser for unlabeled data graphs. Ultimately, logistic regression, laplacian regulariser, and multi-task learning boosted the model's accuracy to 87%. Unlabeled real-world data was utilized to verify the findings [18].

The model for identifying false news that was constructed by Ahmed et al. [19] included six different machine-learning techniques. Six distinct machine learning algorithms, including Decision Tree, Support Vector Machine (SVM), Linear Support Vector Machine (LSVM), Stochastic Gradient Descent, and K-nearest Neighbor, were developed and compared with TF-IDF and Term Frequency (TF) grams such as Unigram, Bigram, Trigram, and Fourgram with a feature size of 1,000, 5,000, 10,000, and 50,000 respectively (KNN). In the evaluation, the Unigram TF-IDF with 50,000 features and the LSVM machine learning algorithm performed remarkably well, obtaining an accuracy of 92% [4]. This particular algorithm has the best accuracy compared to others created to identify false news. You may find an overview of the study on recognizing false news in the following table (Table I).

TABLE I. OVERALL LITERATURE SURVEY IN FAKE NEWS DETECTION

Year	Authors	Method	Accuracy
2015	V.L. Rubin et al.,	presented the logistic regression model	56%
2016	Hai et al.	developed a new semisupervised method for unlabeled data using Laplacian regularized logistic regression	87%
2017	Ruchansky, N., Seo, S., Liu, Y.	Crime Scene Investigation: A Hybrid Deep Model for Identifying Disinformation	89.2%
2017	Granik and Mesyura	A model that is based on the Naive Bayes classifier was proposed.	74%
2017	Gilda	Combining the power of the bi-gram TF-IDF algorithm with the flexibility of the Stochastic Gradient Descent classifier	77.2%
2017	Ahmed et al.,	Implements N-Gram analysis using six distinct AI techniques	92%

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. IMPLEMENTATION

A. Data Acquisition

The first step in developing the models was to collect the necessary data. We must collect reliable information because we want our project to succeed. Kaggle has been an excellent resource for acquiring datasets, but we should remember that we may need to look elsewhere to build a model that will do our project justice. Nonetheless, the dataset has 44,898 entries, each of which has the following four features: title, text, topic, and date. The characteristics of fake news items and actual news pieces may be analyzed and compared using these attributes, which may lead to discovering patterns or trends that distinguish the two types of articles.

The Beautiful Soup algorithm, a Python library that facilitates web scraping of data from websites, is just one example of how we have supplemented our data acquisition process with additional features [13]. Using this algorithm, we could scrape information from primary news sources like CNN, which has proven helpful in NLP research. Vital to our project is the ability to identify fake news; this algorithm has helped us do just that.

For our research, we have looked at news feeds from other websites, like Reddit.com and CNN. As a Python wrapper for the Reddit.com API, we used the Praw API to retrieve information from Reddit.com. Using this API, we got our hands on the most up-to-date and pertinent news stories for our work. As a news aggregation site, Reddit.com has proven invaluable for staying abreast of breaking developments in various fields of study. As a result of using these different strategies for gathering data, we have a more comprehensive set of inputs to use in our analyses, which should lead to more precise results and greater confidence in our project. As researchers, we know how important it is to collect data thoroughly and efficiently before beginning any analysis.

B. Data Preparation

There are multiple steps involved in data preparation that make data more accessible and usable. The first step in processing data is to format it in a data frame. We have used three different sources and several different functions to extract posts. Since data cleaning tasks and analyses can be easily performed on CSV files, we have separated the data into two categories: fake news and real news. In this way, we create a data frame from two different datasets. To guarantee the validity of our analysis, we have used various data preparation strategies.

We cleaned the data by removing duplicates, fixing inconsistencies, and imputing missing values where necessary. This is especially helpful in Word2Vec operations, where two words with the same meaning but different casing may be processed in two different ways. As a next step, we removed duplicates from data frames. Tokenization and lemmatization are two advanced algorithms that allow us to systematically eliminate duplicate words from our data and conduct analyses with a clear and concise vocabulary. A proper word count relies on not counting the same word more than once, so removing duplicates is crucial. When using word2Vec, it is a lot more functional.

We are aware of the potential benefits of incorporating data from social media sites like Reddit into our existing sets of information. We successfully accessed and extracted valuable data from Reddit.com by utilizing APIs like PRAW (Python Reddit API Wrapper) by providing relevant credentials like client secret and user agent in our program. The client secret acts as a key that unlocks the API for our use, while the user agent is a label that gives the Reddit API insight into our project. Using this method, we could pull various information from Reddit.com, such as user-submitted content, comments, and posts. Finally, we have automated our data extraction process with Python scripts, allowing us to scrape Reddit for new data periodically and guaranteeing that our analyses are always based on the most recent data possible. As researchers, we know that various data collection strategies are necessary for producing accurate and insightful results. Our prior success in data mining Reddit has been invaluable in this endeavor.

Using the "requests" and "urlopen" libraries, we run the Beautiful Soup algorithm on CNN.com to scrape the site for helpful information. Using this technique, we can collect massive amounts of textual data for NLP, which will help us refine our analyses and strengthen the credibility of our findings. The first step in utilizing BeautifulSoup with an HTML file for CNN news detection is to import the library and open the file using Python's in-built file handling functions. After the file has been opened, it can be used as an argument in the BeautifulSoup constructor to generate an object that contains all of the document's HTML content.

© 2025 IJMRSET | Volume 8, Issue 8, August 2025| DOI:10.15680/IJMRSET.2025.0808131

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Developers can then use the BeautifulSoup object's methods and attributes to navigate the HTML document and extract data. If you want to find the first occurrence of a specific HTML tag in the document, you can use the find() method, while the find all() method will allow you to find every instance of that tag in the document. Soup.findAll ("p", class ="paragraph inline-placeholder") is used in this project. Locate the paragraph tag that will be our data source for spotting the news.

BeautifulSoup's many uses extend beyond simple tag searching; it can also be used to extract the textual content of HTML elements, extract attributes like class or id, and otherwise manipulate the HTML document by inserting, removing, or rearranging elements.

BeautifulSoup is a highly effective and versatile tool for extracting data from HTML documents. Web scraping and other uses involving HTML data benefit significantly from its user-friendliness and comprehensive documentation.

C. Exploratory Data Analysis (EDA)

By examining and analyzing data to distil their essential features, exploratory data analysis (EDA) is a valuable tool. Exploratory data analysis, or EDA for short, refers to initially reviewing the information to create patterns, identify incongruities, and test predictions using graphical and summary statistical representations. This is a vital step in any data analysis project. As can be seen in Figures 1 and 2, the World Clouds algorithm has been used on datasets including both genuine and fraudulent news.

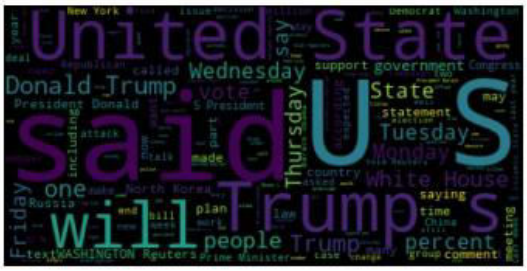

Fig. 1. Word Cloud image of the real news data from Keggle

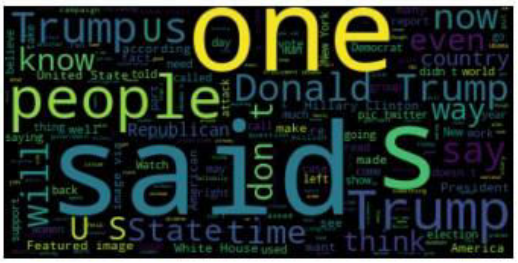

Fig. 2. Word Cloud image of the fake news from Keggle

D. Modeling Phase

All the steps necessary to implement a model trained with machine learning are detailed here. The modelling process consists of four distinct steps. The first step is to merge two existing datasets into a single, final dataset. In the second step, the text is converted into vectors for construction with the help of word2vec. The third step is to divide the information into test and training sets—finally, machine learning model construction.

1). Merging all Datasets

As mentioned, real.csv held true news data, while fake.csv was home to fabricated news files. Merging the real.csv and fake.csv data frames yields the final data frame. Index, title, text, subject, date, source, and label were the only columns remaining in the final data frame. The panda library in the python programming language provides several options for joining multiple data frames into one. In order to create the final data frame for this analysis, we use the append ()

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

function to join together two existing data frames. Since we used the append() function to join the two data frames, everything is in order. True news stays on the top, and fake news stays at the bottom, so we use a function called "shuffle()" to shuffle them. After mixing those data, we finally got all the complicated news for testing and training to remove the bias between these two.

2). Using word2vec

Before running a machine-learning algorithm, we must convert each text into a numerical vector form. We deployed a word2vec system. The text must first be converted into a numerical representation using a tool known as word2vec before a machine learning algorithm can comprehend it. Each matrix represents the words in the dataset. Consequently, the dictionary that contains all of the terms in the dataset defines each document according to the number of times the term occurs in the dictionary. Word2Vec is a well-known technique that may be used to convert word representations into vectors. The neural network-based method gains distributed word representations from big text corpora. Word2Vec's fundamental idea is to put each word in a high-dimensional vector space so that words with related meanings are situated close to one another.

3). Splitting the data into Training and Testing data Training data will account for the vast majority (80%) of the dataset. Using this information, the model can be constructed. The remaining 20% will be used as test data. 80% of the dataset was used to classify, while the remaining 20% was not included in the dataset used to train the model. This percentage (20%) will test how well the model works. With an 80:20 split between training and testing data, Ahmed et al. [19] successfully identified 90% of false news articles. The two data sets will be split 80:20 between training and testing data based on a review of relevant literature.

4). Model building using Sklearn Pipeline

The sklearn pipeline functionality is primarily used to combine multiple estimators into a single one conveniently. This feature shined brightest when the steps required to process data were precisely known. Also, the interface's rich declarative nature simplifies the model's evaluation. Classification, feature selection, and normalization all used this capability in this study. If you want to build this Sklearn pipeline later on, you must predict by calling the relevant function with the testing and training datasets. Adding a classifier was the final step in the sklearn pipeline. As mentioned earlier, a sklearn pipeline was built for each of the different classifiers (models) within their respective n-grams. An n-gram is a series of N words that come after another. An example of a unigram would be "Data," whereas a bigram would be "The Data," and a trigram would be "The Data Science" (trigram). An n-gram has many applications in natural language processing. N-grams are used in various applications, such as spell check and sentence completion. The use of this is widespread in text messages and electronic mail. The machine learning model must be trained using the final dataset, which we have finalized with stopword removal, lemmatization, punction removal, and word2vec. After the machine learning model has been trained, it will automatically classify articles as either real or fake. Statistical Method of Backwards Inference To do this, we first train a logistic regression model on one set of data and then measure its performance on another set of data, called the test set. In addition, this model achieves a success rate of 96.21 percent. Using the joblib library, we create a file to store a trained logistic regression model. The joblib library simplifies the process of storing and retrieving scikit-learn models. Deploying trained models in production environments or sharing models with other researchers and developers are possible with serialized scikit-learn models. The news from Reddit and CNN will both be predicted using this model.

Classification Report and Confusion Matrix:

Test Classification Report:

	precision	recall	f1-score	support
0	0.97	0.94	0.95	1889
1	0.95	0.97	0.96	2204
accuracy			0.96	4093
macro avg	0.96	0.96	0.96	4093
weighted avg	0.96	0.96	0.96	4093

Fig.3.Test Classification Report of Logistic Regression

© 2025 IJMRSET | Volume 8, Issue 8, August 2025| DOI:10.15680/IJMRSET.2025.0808131

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

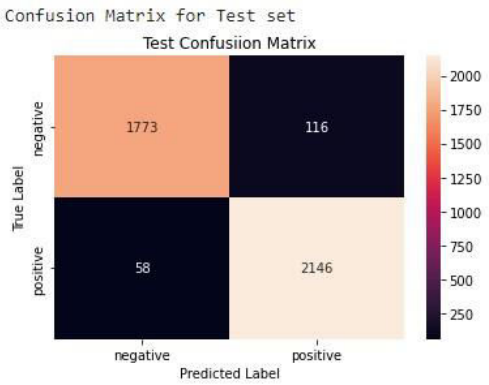

Fig.4. Confusion matrix for the test set of Logistic Regression.

This example shows how to fine-tune an XGBoost classification model's hyperparameters using the GridSearchCV class from sci-kit-learn. With the grid search's assistance, we can systematically test out various hyperparameter settings in search of the one that improves accuracy most dramatically on the training data. In the end, we get 95.93 percent, and we train an XGBoost classification model using the training data and the best hyperparameters discovered via a grid search. In addition, the class weight parameter is set to "balanced" to help correct the mismatch between classes in the training data. As soon as the model has been trained, it can make predictions based on untrained data. Much like in logistic regression, we have archived the training data model for forecasting.

Classification Report and Confusion Matrix:

Test Classification Report:

	precision	recall	f1-score	support
0	0.96	0.93	0.94	1889
1	0.94	0.97	0.95	2204
accuracy			0.95	4093
macro avg	0.95	0.95	0.95	4093
weighted avg	0.95	0.95	0.95	4093

Fig.5. Test Classification Report of XGBoost

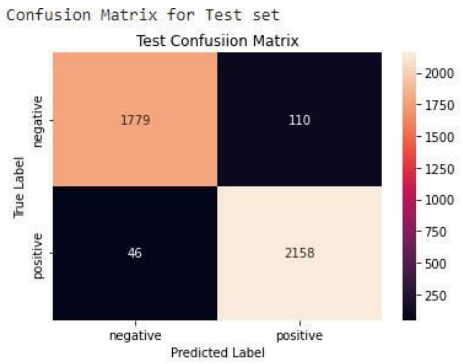

Fig.6. Confusion matrix for the test set of XGBoost.

To determine the optimal hyperparameter for a random forest model, Random Forest uses the GridSearchCV class. With the grid search's assistance, we can systematically test out various hyperparameter settings in search of the one

© 2025 IJMRSET | Volume 8, Issue 8, August 2025| DOI:10.15680/IJMRSET.2025.0808131

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

that improves accuracy most dramatically on the training data. To get this high percentage(94.99%), we run a grid search to find the optimal values for the hyperparameters and use those settings to train a random forest classification model on the training data. In addition, the class weight parameter is set to "balanced" to help correct the mismatch between classes in the training data. Due to the model's versatility, we save time storing it after training is complete; we can then use it to make predictions on unprocessed data.

Report on Classification and a Matrix of Confusion:

Test Classification Report:

	precision	recall	f1-score	support
0	0.96	0.93	0.94	1889
1	0.94	0.97	0.95	2204
accuracy			0.95	4093
macro avg	0.95	0.95	0.95	4093
weighted avg	0.95	0.95	0.95	4093

Fig.7. Classification Report of Random Forest

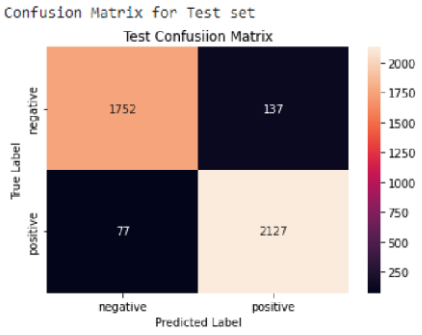

Fig.8. Confusion matrix for the test set of random forest model

E. Prediction Phase

As was discussed earlier in the data preparation phase, the final step involves retrieving data from online sources like CNN and Reddit. Assuming we have the information we need, we can load the three trained models we created to foretell the news's outcome. To detect CNN news in our project, we need only copy the URL from the webpage and paste it where "url=" appears in our project. In the next step, the text from that HTML file will be extracted because we used beautifulsoup to find the paragraph tag and place it in a new data frame. Then, we will start the data preparation process as we have mentioned in step B. Moreover, The three models, as mentioned above, are then loaded to make an accurate forecast.

	text	RF_p	XGB_pred	logreg_p
0	While the bill would not set a national ...	1	1	1
1	The push for a vote on federal legislati...	1	1	1
2	The bill, called the Respect for Marriag...	1	1	1
3	In the event the Supreme Court might ove...	1	1	1
4	House Speaker Nancy Pelosi wrote in an o...	0	1	1
5	"Just as I began my career fighting for ...	0	0	0
6	President Joe Biden applauded Senate pas...	1	1	0
7	Biden added: "I look forward to welcomin...	1	1	1
8	The Supreme Court's move in June holding...	1	0	1
9	The Supreme Court is currently consideri...	1	1	1
10	Several conservative members of the Supr...	0	0	0

Fig.9. The prediction result of the CNN website by using 3 models.

© 2025 IJMRSET | Volume 8, Issue 8, August 2025| DOI:10.15680/IJMRSET.2025.0808131

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

We do the same thing with Reddit, only this time using the site's API to collect data. Only political news stories (of 400 total) are included in the data. Furthermore, we save that news to our new data frame. After collecting information, we clean it up by doing things like lemmatizing it and taking out punctuation so and so far. Then we can predict that news from Reddit. We then predict the outcome after loading those three models.

	Title	RF_pred	XGB_pred	LR_pred
0	demands trump removal grow fascist speech cond...	0	0	0
1	democrats prepare bill limiting us supreme cou...	1	1	0
2	proud boys intended kill mike pence nancy pelo...	0	0	0
3	white house ordered nih cancel coronavirus res...	1	1	1
4	joe bidens gettysburg speech wa 25 minute long...	0	0	1
...	...	...	...	...
295	republicans voted feeding baby they're forcing	0	0	0
296	pence break silence take credit pfizer vaccine...	0	0	0
297	cry river sanders hits back billionaire invest...	0	0	1
298	americas top military officer say take oath king	0	0	0
299	man foot desk pelosis office capitol arrested	0	0	0

Fig.10. The prediction result of Reddit News Content by using 3 models.

VII. RESULTS

The results of our three machine learning models for detecting false news are shown in Table II. Regarding recognizing fake news, the Logistic Regression model performs far better than its rivals. By combining their resources into a single strategy, known as an ensemble method, all of the separate models see improvements in their performance.

TABLE II RESULT OF THE THREE DIFFERENT MODELS

Model	Accuracy
Model 1: Logistic Regression	96.21%
Model 2: XGBoost	95.93%
Model 3: Random Forest	94.99%

Finding the best model for accurately predicting fake news was a primary motivation for this study. However, the extent to which the result can be trusted will depend heavily on the best model's performance. Table II's Accuracy shows that Logistic Regression is the most effective model for identifying false news articles, with a 96.21% success rate. This study approaches the problem of identifying fake news as a yes/no question in which a post is either authentic or not.

The highest accuracy attained by Logistic Regression was 96.21%. One common way to summarize a classification model's results is with a straightforward table called a confusion matrix. The model's confusion matrix is depicted in Fig. 8.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

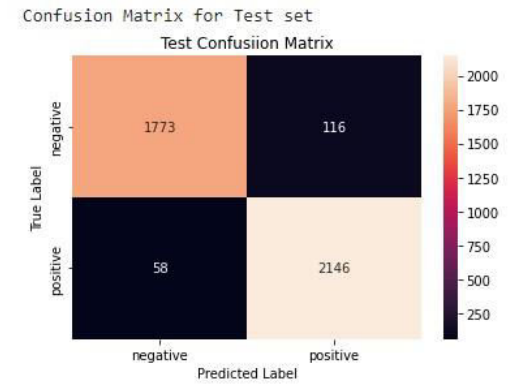

Fig.11. Confusion matrix of the best model

VIII. CONCLUSION

To that end, this study compares three approaches to identifying fake news in Reddit threads. The Logistic Regression was the most effective among the models tested for identifying fake news in Reddit threads and CNN articles. The distinguishing features of the three models are as follows:

1. Model type: In the form of a linear model known as logistic regression, the sigmoid function is used to derive a probability from a data collection. XGBoost is an ensemble model, much like Random Forest, except instead of using a single tree to generate predictions, it employs a network of linked nodes.
2. Training method: Batch gradient descent is used to train Logistic Regression and Random Forest, decision tree induction is used to train Random Forest, and XGBoost is trained with gradient boosting.
3. Handling non-linear relationships: For its analysis, Logistic Regression makes the linear assumption that the predictor and the target variables are related in some fashion. In addition, both XGBoost and Random Forest are adaptable to situations where the predictor and outcome variables are not linearly related.
4. Interpretability: Logistic Regression produces coefficients that may be used to describe how each predictor variable affects the target variable, making it more comprehensible than XGBoost and Random Forest. XGBoost and Random Forest are considered "black box" models because the results can be more challenging to interpret.
5. Handling missing values: While Random Forest cannot deal with missing values, Logistic Regression and XGBoost can.
6. Handling class imbalance: XGBoost and Random Forest can handle class imbalance without under-sampling or oversampling by employing class weights or SMOTE, but Logistic Regression may need such adjustments.

Overall, each algorithm has its pitfalls and benefits; picking one will depend on the problem's nature, the available data size, and the project's goals.

REFERENCES

[1] Ankit Kumar P, and Kevin M, "Fake News Detection on Reddit Utilising CountVectorizer and Term Frequency-Inverse Document Frequency with Logistic Regression, MultinominalNB and Support Vector Machine", 2021, 32nd Irish Signals and Systems Conference (ISSC).
[2] Hunt A, and Matthew G, "Social Media and Fake News in the 2016 Election", Working Paper 23089, Working Paper Series, National Bureau of Economic Research, Jan 2017
[3] Y. Zhang, Y. Su, L. Weigang, and H. Liu, "Rumor and authoritative information propagation model considering super spreading in complex social networks," Physica A: Statistical Mechanics and its Applications, vol. 506, no. Physica A 499 (2018) 276-287, pp. 395– 411.
[4] Başţanlar, Y., Ozuysal, M. (2014) "Introduction to machine learning", Methods in Molecular Biology (Clifton, N.J.), 1107, 105–128
[5] Oladipupo, T. (2010) "Types of Machine Learning Algorithms", N e w A dv a n ce i n M a c h i n e L e a r n i n g, University of Portsmouth United Kingdom.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[6] Jones, M.T. (2017) "Models for machine learning", IBM Developer.
[7] Portugal, I., Alencar, P., Cowan, D. (2018) "The use of machine learning algorithms in recommender systems: A systematic review", Expert Systems with Applications, 97, 205–227.
[8] Abdullah-All-Tanvir, Mahir, E.M., Akhter, S., Huq, M.R. (2019) "Detecting Fake News using Machine Learning and Deep Learning Algorithms", in 2019 7th International Conference on Smart Computing Communications (ICSCC), Presented at the 2019 7th International Conference on Smart Computing Communications (ICSCC), 1–5.
[9] Aborisade, O., Anwar, M. (2018) "Classification for Authorship of Tweets by Comparing Logistic Regression and Naive Bayes Classifiers", in 2018 IEEE International Conference on Information Reuse and Integration (IRI), Presented at the 2018 IEEE International Conference on Information Reuse and Integration (IRI), 269–276.
[10]Tompkins, J. (n.d.) "Disinformation Detection: A review of linguistic feature selection and classification models in news veracity assessments", Oct 26 2019.
[11]Manning, C., Schutze, H. (1999) "Foundations of Statistical Natural Language Processing", MIT Press.
[12]Silva, C., Ribeiro, B. (2003) 'The importance of stop word removal on recall values in text categorization", Proceedings of the International Joint Conference on Neural Networks, 2003., Neural Networks, 2003. Proceedings of the International Joint Conference on, Neural networks, 3, 1661.
[13]Al Asaad, B., Erascu, M. (2018) "A Tool for Fake News Detection".
[14]Kaur, J., Saini, J.R. (2015) "A Natural Language Processing Approach for Identification of Stop Words in Punjabi Language".
[15]Balakrishnan, Vimala and Lloyd-Yemoh, Ethel (2014) "Stemming and lemmatization: A comparison of retrieval performance". in: Proceedings of SCEI Seoul Conferences, 1011 Apr 2014, Seoul, Korea.
[16]Rubin, V.L., Chen, Y., Conroy, N.J. (2015) "Deception detection for news: three types of fakes", in Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, ASIST '15, American Society for Information Science: St. Louis, Missouri, 1–4.
[17]Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H. (2018) "FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media", 12.
[18]Hai, Z., Zhao, P., Cheng, P., Yang, P., Li, X.-L., Li, G. (2016) "Deceptive Review Spam Detection via Exploiting Task Relatedness and Unlabeled Data", in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Presented at the EMNLP 2016, Association for Computational Linguistics: Austin, Texas, 1817–1826.
[19]Ahmed, H., Traore, I., Saad, S. (2017) "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques", in Traore, I., Woungang, I. and Awad, A., eds., Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, Lecture Notes in Computer Science, Springer International Publishing: Cham, 127–138.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com