Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

ISSN NO. 2320-5407

Article DOI: 10.21474/IJAR01/19332
DOI URL: http://dx.doi.org/10.21474/IJAR01/19332

RESEARCH ARTICLE

DETECTION OF FAKE NEWS ON TWITTER USING MACHINE LEARNING: AN XGBOOST-BASED APPROACH WITH SENTIMENT AND SOURCE CHARACTERISTIC ANALYSIS

Dhvani Gupta
Calcutta International School, 724, Eastern Metropolitan Bypass Rd, Near Fortis Hospital, Anandapur, VIP Nagar, Kolkata, West Bengal 700107, India.

……………………………………………………………………………………………………....

Manuscript Info

……………………..

Manuscript History
Received: 21 June 2024
Final Accepted: 24 July 2024
Published: August 2024

Key words:-
Fake News, Social Media, Twitter, Machine Learning, Ensemble Learning, Sentiment Analysis, Detection Algorithm, XG Boost, Truthseeker 2023 Dataset

Abstract

………………………………………………………………

The spread of fake news on social media platforms is becoming an increasingly alarming problem with fake news becoming more deceptive and harder to detect. Twitter, in particular, poses a significant threat as fake news spreads faster than real news on the platform, enhancing misinformation and leading to serious consequences.This project presents a novel machine learning-based approach for detecting fake news tweets on Twitter using the TruthSeeker 2023 dataset from the University of New Brunswick. As the largest ground truth dataset for fake news detection on social media, it contains over 130,000 crowdsourced tweets, enabling the creation of a broader and more applicable model for real-world scenarios. The algorithm employed in this study leverages the properties of gradient-boosted decision tree models (XGBoost) to develop a novel method for classifying fake and real news tweets. The proposed model preprocesses the data by extracting additional features for each tweet, such as detailed sentiment analysis of both the tweet and the related news statement, as well as features pertaining to the author. These features are added to the tweet's feature vector. The enhanced feature vectors are then fed into an XGBoost model with tuned hyperparameters determined through a grid search algorithm to perform binary classification. The additional extracted features increase the robustness of the model by highlighting key differentiating factors between real and fake tweets. The results of this study demonstrate the effectiveness of the proposed algorithm, achieving an accuracy of 0.9335 on over 13,000 unseen tweets.

……………………………………………………………………………………………………....

Introduction:-

The web provides a highly interconnected worldwide platform for the interchange and spread of information amongst all its users, within a matter of few minutes at little to no cost. This phenomenon has led to the increase of accessibility to real time citizen journalism, as well as improved visibility of both real and fake news, establishing social media platforms such as Twitter as a main conduit of digital news. In recent years, the spread of misinformation on social media has reached alarming levels. A single false tweet can now travel to millions of users in mere minutes, potentially causing widespread harm.

Corresponding Author:- Dhvani Gupta
Address:- Calcutta International School, 724, Eastern Metropolitan Bypass Rd, Near Fortis Hospital, Anandapur, VIP Nagar, Kolkata, West Bengal 700107, India.

Research Background

This section provides a foundational understanding of fake news, exploring its significance in fake news detection in social media.

What is fake news?

The web provides a highly interconnected worldwide platform for the interchange and spread of information amongst all its users, within a matter of few minutes at little to no cost. This phenomenon has led to the increase of accessibility to real time citizen journalism, as well as improved visibility of both real and fake news, establishing social media platforms such as Twitter as a main conduit of digital news [1]. Fake news refers to false information that is created and spread on social media and collaborative platforms with an intent to influence political opinion [2] and earn money [3], and threatens all aspects of human society including healthcare, economy, and morality. Fake news is characterized by deliberately misleading information, lack of reliable references, and intentional ambiguity, leading to misinformation and disinformation.

What was the advent of fake news?

The spread of misinformation and disinformation is not new and cannot be attributed solely to social media; it has been used historically to sway national ideology. Much before the advent of social media, false news stories spread by The Associated Press helped lead to the inauguration of Rutherford B. Hayes as president and the end of post-Civil War Reconstruction in the United States of America. “yellow journalism,” sensational coverage sold newspapers at the expense of factual accuracy at the turn of the 20th century, which strongly contributed to the start of the Spanish-American War and (arguably) the U.S. entry into World War I [4]. Thus, false news has historically been used as a weapon to create propaganda, and induce anxiety and fear amongst the unsuspecting public, the effects of which have been amplified since the advent of the internet and social media.

When did fake news start spreading on Twitter ?

The 2016 U.S. Presidential election marked a significant rise in fake news on social media, particularly Twitter, where fabricated stories appeared credible.Two of the biggest false hits of the incident were a story claiming Clinton sold weapons to ISIS and a hoax claiming the pope endorsed Trump were shared hundreds of thousands of times on the eve of the election. The incident also revealed a key characteristic of the spread and properties of fake news in social media where fake news accounted for nearly 6% of all news consumption, but it was heavily concentrated—only 1% of users were exposed to 80% of fake news, and 0.1% of users were responsible for sharing 80% of fake news [2]. Since 2016, the spread of fake news has become more widespread relating to incidents such the covid 19 pandemic [5] and climate change [6] becoming a significant area of concern.

How does fake news spread ?

Humans are susceptible to the spread of false information, however the creation and spread of fake news is attributed largely to the nefarious actors and bots on the internet which may act independently or as a part of a greater social network. They are used to deceive readers to create an illusion of consensus towards a piece of news by echoing it many times and or expressing direct support for it, thus used to artificially engineer virality and spread fake news [7]. Lone-wolves create multiple fake accounts to coordinate similar reviews or comments, making their views appear more widespread than they are. Lone-wolf operations using multiple accounts, which commonly hold deceptive intentions [8] and central locations in social networks to avoid detection, can be especially convincing as readers are typically not aware that a whole discussion is fabricated and originates from a single source. On a larger scale, social media botnets are used to spread false information. Bots, which are fake or compromised accounts controlled by a single individual or a program, serve two main purposes: to share the same information with a large audience quickly, and to inflate the “social status” of certain users, both of which make false information appear credible and legitimate [9]. In a recent study by Vosoughi et.al [10], analysing over 126000 false information cascades on Twitter revealed that humans played a larger role than bots in the spread of misinformation in twitter due to retweeting of false information by non- bot accounts. with regard to the tweeters, some of them play their role in such a way that soon turns them into the epicentre of diffusion (sometimes known as ‘Super Spreaders’), while on the other hand there will be many tweeters of the same content that find no success to have their tweets retweeted by anyone [11].

What is the rationale behind deception through fake news?

A study by Vosoughi et.al [10] found that false information spreads faster and deeper on Twitter than true information. This is primarily due to human susceptibility to false information and the difficulty in distinguishing between real and fake news. For fake news, a recent study by Pérez-Rosas et.al [12] created a dataset of crowdsourced and crawled celebrity-oriented real and fake news, and gave 680 pieces of news (50% fake) to two humans to identify fake ones from them, who achieved an average marginal accuracy of 70.5% in detecting made-up crowdsourced news, and 78.5% in detecting celebrity news. The uprise of improved recommendation and algorithms and personalized individual suggestions based on one's interest has led to the creation of 'echo chambers' on social media, which creates a polarizing effect on content and amplifies the spread of false information by lowering the bar for critical thinking and increasing exposure to one kind of false content thus increasing its perceived accuracy [13]. Moreover, false information may spread due to leverage of preexisting bias, confirmation bias [14], and misinterpretation of information creating misinformation.

What are the types of fake news ?

False news can be categorized as misinformation or disinformation based on the intent to deceive. Misinformation, which lacks deceptive intent, often arises from misinterpretation, and can cause anxiety and panic, as seen on Twitter [15]. Disinformation [16] however is created with the intent to deceive and spread deliberately fabricated false information to sway public opinion may prove to be more malicious [7], as during the 2016 incident on twitter. False news may further be classified on twitter as false opinions where there is no existence of a ground truth and used to influence the readers opinion such as on e-commerce review platforms or fact-based false news which comprises fabricated information which directly contradicts a ground truth fact.

What is the impact of fake news ?

Fake news has significant real-world impacts, including stock market fluctuations, increased terrorist activity, and higher vaccine refusal rates [17]. Social media impact of fake news is measured as engagement with fake news articles through statistics such as amount of time survived without being removed, number of times itself or related tweets were shared and retweeted, and the number of people it reached. The recent study by Vosoughi et.al [10] found that fake news spread faster (more retweets within a particular time frame) ,deeper, and broader(reaching a wider network of audience through retweets). The top 1% of false tweets reached over 1,000 users, which true information tweets rarely did and false information was six times faster in reaching the same number of people as true information. Thus, the spread of false information, though highly concentrated (top 1%), may prove to be highly impactful and persuasive, by spreading widely and quickly across the web.

Related Work

There is a vast literature which studies the spread of false news on social media. This section provides an overview of some recent research contributions in this space.

Algorithms developed to detect fact-based fake news such as hoaxes, fake news, and rumours on social media can be categorized into two major approaches: feature engineering-based methods and propagation-based methods.

Feature Engineering-Based Methods

These methods have been employed to identify various types of malicious users and activities, such as bots, trolls, vandals, and sockpuppets. They rely on various features derived from the text, user information, network properties, and metadata to distinguish true information from false. Text-based features are foundational, often categorized into stylometric (e.g., word length), psycholinguistic (e.g., LIWC), and complexity-oriented (e.g., readability indices). The general approach involves extracting these features and using them in machine learning models to classify information. Studies like those by Qazninian et al. [18] and Gupta et al. [19] demonstrate high accuracy using content features such as unigrams, bigrams, and part-of-speech tags, complemented by user and Twitter-specific features (such as hashtags and URLs). Recent research by Perez-Rosas et al. [12] and Horne and Adali [20] has shown that while text-based features are effective, their performance can decline as adversaries adapt, emphasizing the need to incorporate a broader set of features. For instance, combining text, user, network, and metadata features has been shown to significantly improve detection performance, achieving high accuracy (typically in the 80s and 90s) and practical utility.

Propagation-Based methods

These methods examine the spread of information across social networks to identify misinformation. Propagation models simulate the spread of information, focusing on how false information propagates differently compared to true information. Propagation-based models serve two primary purposes: simulating the spread of false information and developing strategies to mitigate its impact. For example, models like SEIZ categorize nodes into states (susceptible, exposed, infected, and sceptical) and use these states to predict the spread of false information. Incorporating propagation information into machine learning frameworks has also been effective, with studies showing improved accuracy when features like the number of replies and retweets are included. Mitigation strategies involve spreading true information to counter false narratives, significantly reducing the lifetime and reach of rumours. Research by Acemoglu et al. [21], Jin et al. [22], and Tripathy et al. [23] highlights the effectiveness of these methods in both detecting and mitigating the spread of misinformation.

These detection methods, whether feature-based or propagation-based, have proven to be effective in identifying and combating misinformation in their respective datasets to identify false information, and usually achieve a high accuracy, precision, or AUC score in the 80s or 90s.

Challenges of Fake News Detection

Despite numerous efforts to curb this issue, false information continues to proliferate various platforms, influencing public opinion and decision-making. The central challenge lies in accurately identifying and mitigating the spread of such misinformation. The accurate detection of false information on social media platforms poses a few key challenges which still make this problem difficult to address. The main challenge in detecting false news on social media lies in the sheer volume and rapid speed at which false information spreads online making it difficult for detection systems to respond instantly, accompanied by its sophisticated, deceptive, and evolving nature, diverse sources and formats, and the complex social dynamics that amplify its spread and further complicate detection efforts. Additionally, the variability in language and cultural context, ambiguity in content due to lack of appropriate context, limited availability of labelled data for training algorithms, privacy and ethical concerns, and the need for effective coordination to prevent spread of misinformation across different platforms make addressing this challenge even more complex.

Proposed solution

Previous research has primarily focused on both text-based and propagation-based detection methods to distinguish between real and fake news tweets on twitter, by using content features extracted from the tweets such as parts-of-speech tags, sentence lengths, etc. and network features such as the number of retweets, hashtags and followers of the user who posted the tweet. However, there is a lack of comprehensive studies that utilize these feature-based detection methods to draw comparisons between the tweets and the news articles they relate to, or explore its possible applications in fake news detection. Motivated by the limitations of current approaches, this paper aims to develop a machine learning model to apply feature-based detection methods in examining the features of both the tweet itself and the news articles it relates to, and analysing the relationship between the two to investigate its effectiveness in differentiating between real and fake news tweets. Specifically, it investigates how extracting certain additional features from the tweet and the related articles can improve detection accuracy of the chosen model when compared to a similar model trained solely on the features of the tweet. This research is significant because it offers a more robust and novel approach to fake news detection on twitter.

Methodology:-

This section explores the Truthseeker 2023 dataset utilized for model training, provides insightful data visualizations illustrating key differentiating features between real and fake news, and introduces the detailed architecture of the XGBoost model.

Dataset

The Truthseeker 2023 is a dataset of tweets and tweet features that contains both real and fake news tweets [24]. It is the largest ground fake news analysis dataset for real and fake news content in relation to social media posts. The data for the Truth Seeker and Basic ML dataset were generated through the crawling of tweets related to Real and Fake news from the PolitiFact Dataset. Taking the ground truth values for each tweet and the crawling for tweets related to the news topics (using manually generated keywords to input into the twitter API), over 133,000 tweets related to 579 real and 479 fake pieces of news were extracted.

This dataset is beneficial to my fake news detection task because it contains real-world examples of tweets by actual users that provide more realistic and accurate results than synthetic datasets. The CIC Truthseeker 2023 dataset has several features that are helpful for our analysis, including the inclusion of real and fake news tweets and information related to the related news article headline and author of the statement. The Canadian Institute of Cybersecurity has also extracted multiple features as described in 2.1.1 that are used to visualize the dataset in a

CSV format. While this research aims to make our model as dataset-agnostic as possible, this will allow us to create a comprehensive frame of reference to visualize and train/test on.

Table 1:- Contains some basic statistics related to the dataset.

Statistic Field	**Number**
Total Unique Tweets	134,198
Total real tweets	68,985
Total false tweets	65,213
Fraction of Real tweets	0.514
Fraction of False Tweets	0.486
Total Unique Statements	1,058
Total real statements	579
Total false statements	479
Total authors	161

Table 1: Key statistics for the CIC Truthseeker 2023 dataset

Dataset Features

As shown in the detailed analysis (see Appendix), the dataset includes various features that were essential for training the model. A few of the features of the dataset are classified and summarized below. **Tweet Features** - contains features such as the tweet itself, the news article headline, name of the author and the manual keywords which were used to search the twitter API.

1. **Text features** - contains textual features about the tweet such as number of unique words, average word length, and total word count.
2. **Lexical features** - contains features extracted from the characters of the tweet itself such as number and type of verb and verb tenses, punctuations, and word frequency.
3. **Metadata features** - contains the attributes of the user who posted the tweet including number of followers and friends, retweets, hashtags, and credibility scores for each user.
4. **Additional Extracted Features** - Contains features extracted using pre-existing features in the dataset. The fraction of long words per tweet was extracted using the ‘number of long words’ and ‘total word count’ features from the ‘Features for the Traditional Machine Learning model’ dataset and is calculated as the fraction of long words per the total word count for each tweet in the dataset. The sentiment of the tweet and sentiment of the statement was extracted using the NLTK Library’s pre-trained Sentiment Intensity Analyzer to obtain the negative, positive, neutral, and compound score for each tweet and each unique statement. The sentiment analyser was run multiple times to obtain multiple sentiment scores for each tweet and statement to calculate the final mean sentiment score (calculated as an average of the outcomes of each individual run), to obtain a more accurate result. The compound scores were used to calculate the difference between the compound sentiment of the statement and the average compound sentiment for the tweets correlating to each statement to visualize the variation in sentiment of the tweets from the original news headline for both true and fake news. The average sentiment per author calculated as the mean sentiment compound score of the statement for each unique author, the average statement length per author, the number of unique statements per author, and the fraction of true statements per author were also extracted to characterize certain features to the authors of true and fake news.

Refer to Appendix for detailed description of the features in the dataset.

Data Visualisations

This section employs data visualizations to enhance our exploration of distinguishing features between real and fake news tweets. These visual representations play a crucial role in illustrating patterns and relationships within the dataset, providing clear insights into the characteristics that differentiate genuine information from misinformation on social media.

Credibility score distribution for users.

Figure 1 shows the empirical distribution curve for the credibility score assigned to each user who posted the tweet for both real and fake news tweets. The blue curve shows the distribution of the credibility score for users who posted the real tweets, and the orange curve represents the distribution of the credibility score for the users who posted fake news tweets. The graph shows that the users who posted the fake tweets have a lower average credibility score rating than those who posted the fake tweets.

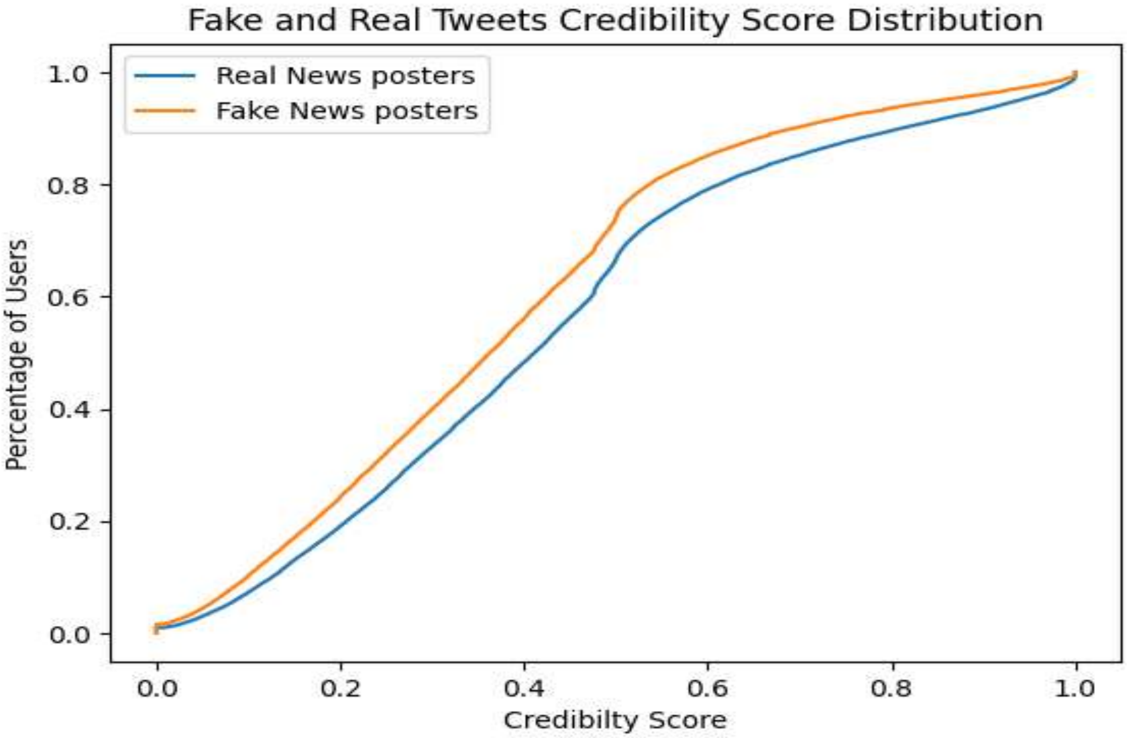

Figure 1:- Distribution of the credibility scores assigned to users who posted tweets, comparing those who posted real news (blue) versus fake news (orange). This graph shows that users who posted fake tweets generally have lower credibility scores.

Distribution of the difference between compound sentiment score for the tweet and the corresponding statement.

Figure 2 shows the empirical distribution for the difference between the compound sentiment score of each unique statement and the average compound sentiment score of each of the tweets associated with that statement. The blue curve shows the distribution for the tweets with the true statement , and the orange curve represents the distribution for the news tweets with the false statements. The sigmoid function for the real statements is deviated farther from the mean, indicating a more even sentiment distribution and greater deviation from the sentiment of the statements than the tweets for the fake statements. Thus, from the data, it may be inferred that the users tweeting about real news had a more varied response ,sentiment wise, from the news headlines when compared to those tweeting about the fake news, whose response was more concentrated, less varied, and more aligned with the sentiment of the news headlines than the real tweets.

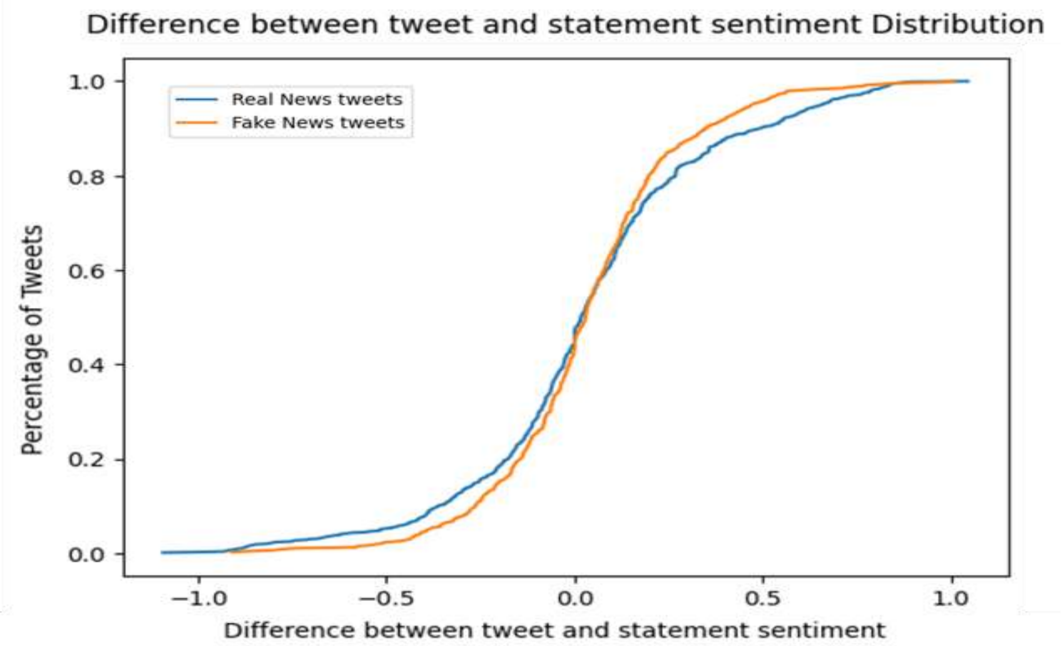

Figure 2:- Empirical distribution of the difference between the compound sentiment scores of tweets and their corresponding statements. The distribution for tweets with true statements (blue) shows a more varied sentiment response compared to tweets with false statements (orange).

Unique Word Count Distribution in Real and Fake tweets.

Figure 3 shows the distribution of the unique word count in each tweet. From the histogram it may be inferred that the real tweets had, on average, a greater unique word count than the tweets relating to fake news. Thus, much like fake news headlines [20], tweets relating to fake news are shorter and more repetitive to catch the attention of the readers and to promote a greater retweet rate for fake news.

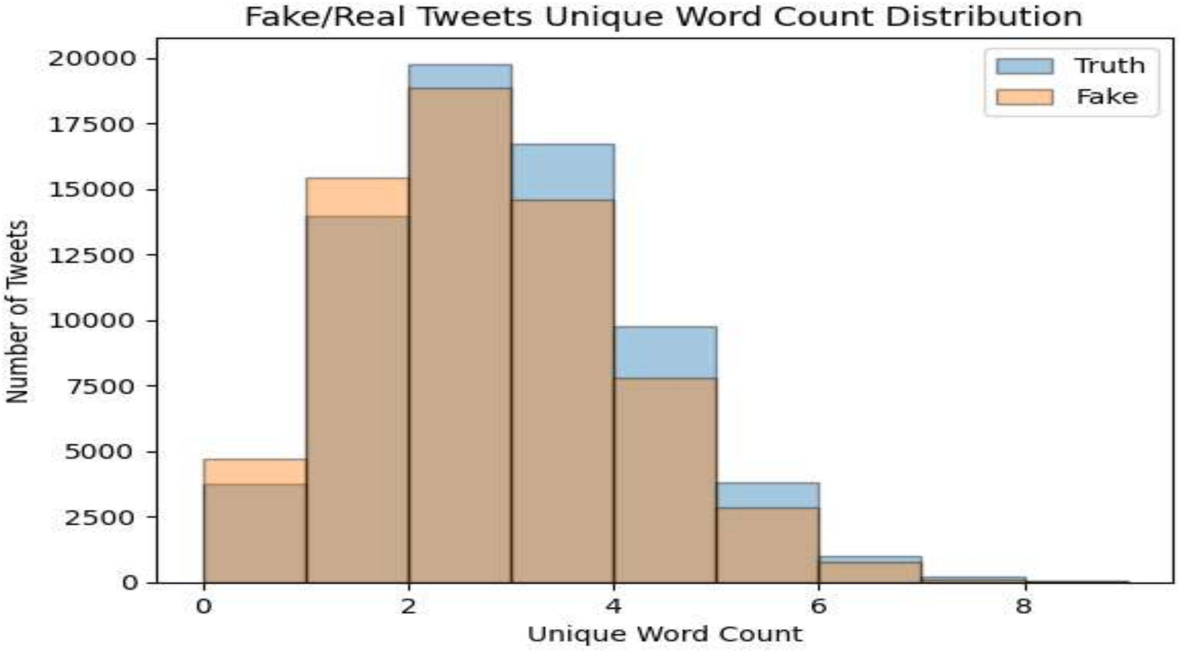

Figure 3:- Histogram of unique word counts in tweets. The data indicates that tweets about real news generally contain a higher unique word count compared to tweets about fake news, which tend to be shorter and more repetitive.

Follower Count Distribution for users.

Figure 4 shows the distribution of the follower count for each user who tweeted the tweets for both fake and real news. The follower count for the accounts posting the fake news can be seen to be significantly lower than those who posted the real news. This statistic may indicate the presence of ‘throwaway accounts or bots which are newly registered accounts with low writing experience and low credibility in the propagation of fake news on twitter.

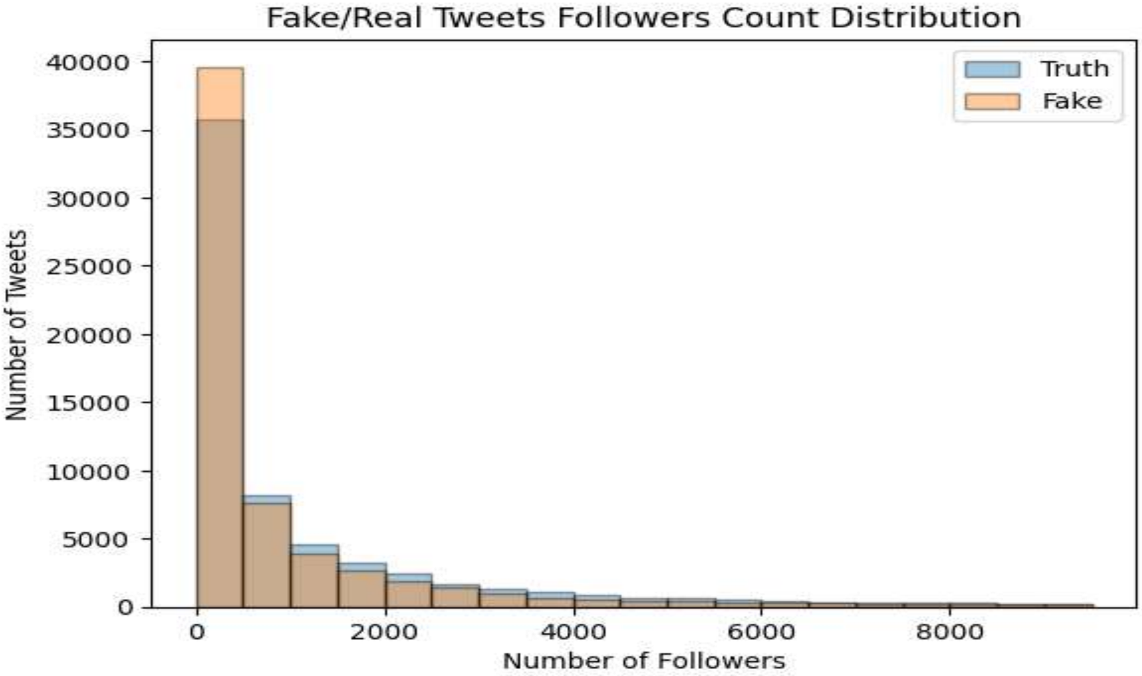

Figure 4:- Distribution of follower counts for users who posted real news versus fake news. Users who posted fake news tend to have significantly fewer followers, suggesting the presence of 'throwaway accounts' or bots.

XGBoost Model Architecture

The machine learning model employed here classifies the tweets as either real or fake news based on the labelled examples of tweets in the training data. The model used for this purpose is a gradient boosted, decision tree-based model known as XGBoost. XGBoost is an ensemble learning method which offers a systematic solution to combine the predictive power of multiple learners, resulting in a single model which gives the aggregated output from several base models. This model uses a technique called ‘boosting,’ where each decision tree or base model is built sequentially such that each subsequent tree minimizes the error , or loss, of the previous tree. The base models, also known as ‘weak learners,’ are decision trees with usually only one split called a decision stump, having high bias and low predictive power. Each of the weak learners contributes some vital information for prediction, enabling the boosting technique to produce a strong learner, or boosted ensemble, by effectively combining these weak learners.

Taking the example of a boosted ensemble comprising only two weak learners (f(1) and f(2)), the model may be visualized as

Boosted ensemble = First Tree (f(1)) + η * second tree (f(2))
Such that, Loss (Boosted Ensemble) < Loss (First Tree)

Where η is the learning rate which is a parameter that controls the step size at which the algorithm makes updates to the model weights. The Loss function is a predefined function that takes the predicted value by the model and the target value as the inputs to quantify the error margin between the two.

Since the objective is to minimize the loss function with each subsequent weak learner, the second tree is fit on the negative derivative of the loss function with respect to the output of the first tree, also known as the ‘residual’ value.

Second tree (f(2)) =$f\left(\frac{-\partial\ Loss\ Function\ \ (output\ \ f(1), target\ \ data\ )}{f\partial\ output\ \ f(1)}\right.$, dataset features)

Thus, for any step M, gradient boosting produces a model such that ensemble at step m equals the ensemble at step M -1 plus the learning rate times the weak learner at step M.

f(M) = f(M - 1) + η * f($\frac{-\partial\ Loss\ Function\ \ (output\ \ f(M-1), target\ \ data\ )}{f\partial\ output\ \ f(M-1)}$, dataset features)

f(M) ≈ f(M - 1) + η * $\frac{-\partial\ Loss\ Function\ \ (output\ \ f(M-1), target\ \ data}{f\partial\ output\ \ f(M-1)}$,

The resultant model is a summation of the result of all models from 1 to n such that n is the number of specified estimators, or weak learners in the model.

F(x) =$\sum_{k=1}^{n} f(k)$

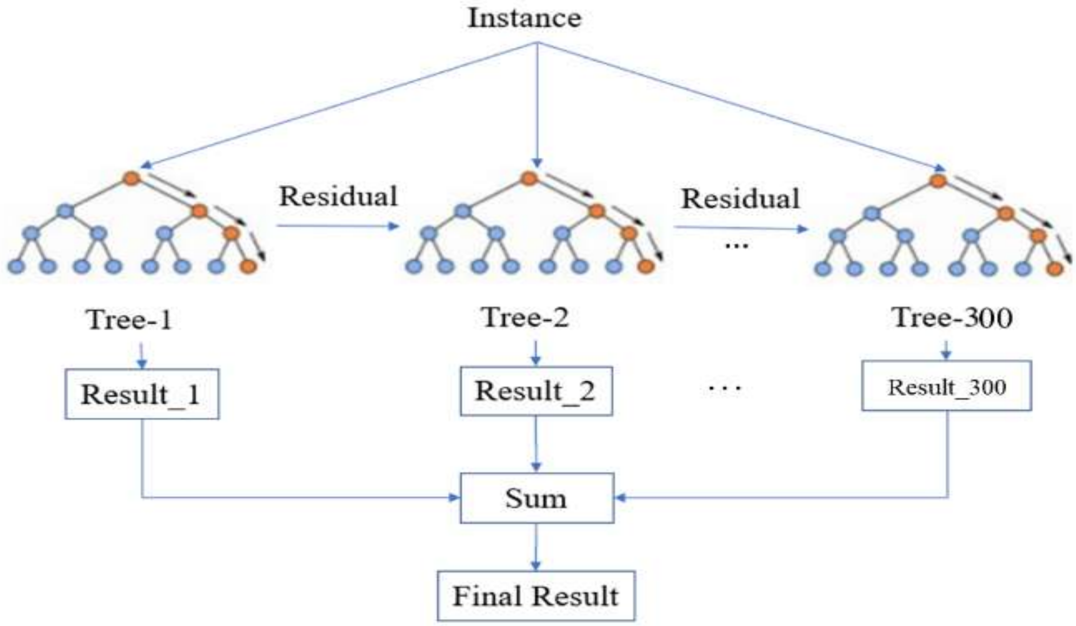

Figure 5:- XGBoost Model Architecture used for tweet detection [25]

Experimental Findings

This section comprehensively outlines the training and evaluation procedures for my XGBoost model, accompanied by a report of commonly used evaluation metrics for measuring the performance of the model. Supervised learning serves as the foundation of my methodology, leveraging labelled datasets where each tweet is distinctly categorized as either fake news or real news tweet.

Training

The model was trained on the CIC Truthseeker Dataset 2023, as discussed in 2.1. As per convention the dataset is split into a training, validation, and testing distribution of 80 : 10 : 10 respectively. Table 2 depicts the distributions for each dataset after the train-test-validation split. The inclusion of a validation dataset allows the model to tune its hyperparameters that fine tune the predictions of the model. The validation dataset also helps the model increase its robustness and generalize effectively to prevent it from overfitting, which is a common issue that arises with the use of XGBoost models because of their complex and fine-tuned nature. By testing the model against a separate unseen test dataset, different from the validation dataset used in the boosting algorithm of the model, the model's performance can be assessed to check for overfitting.

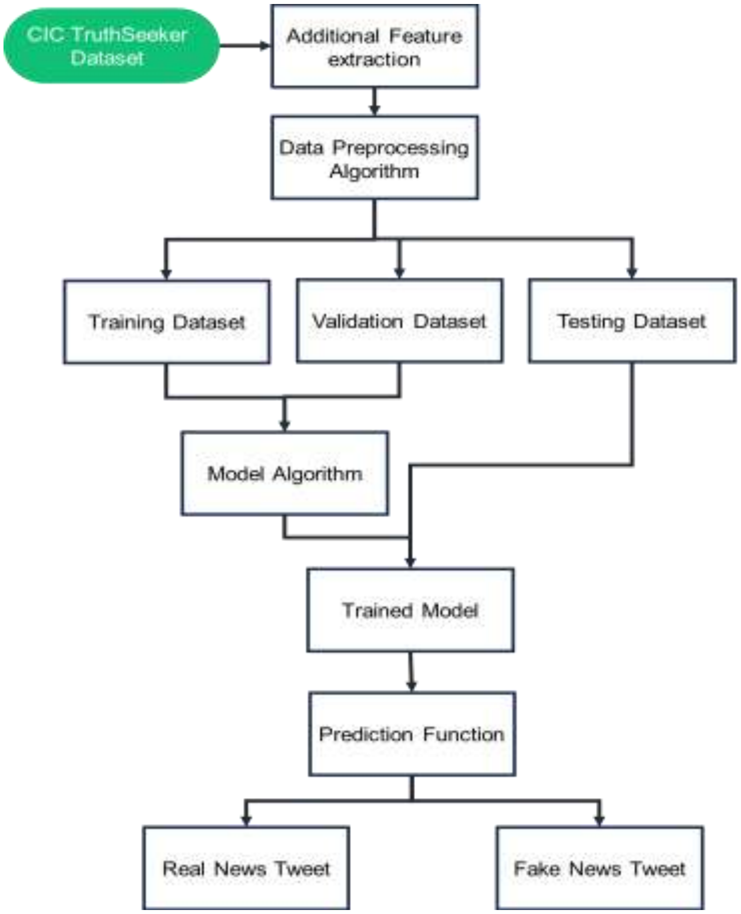

Figure 6:- Illustration of proposed procedure.

Table 2:- Dataset Distributions.

Dataset	**Number of Samples**
Training Set	107358
Validation Set	13420
Test Set	13420
Total	**134198**

For optimization during training the log-loss function, a loss function for binary classification problems, is employed as the predetermined loss function to tune the model. Training and evaluation occur on the pre-processed dataset, utilizing the loss function described above. The inclusion of a validation dataset consistently reduced loss over each weak learner, while tuning the dataset to perform well on the test data without overfitting. A maximum of 100 rounds was permitted for each learner, where training halts if no discernible improvement in loss minimization is observed for 100 consecutive rounds.

The training process involved grid search cross-validation to perform hyperparameter tuning for the number of estimators, maximum depth, and learning rate of the model. The hyperparameter pairings were optimized by testing each iteration using the performance metric of Area under the ROC curve.

Common Performance Metrics

To evaluate the performance of my XGBoost model in the realm of fake news detection, severalwell-established performance metrics are utilized. These metrics provide invaluable insights into the model's capacity to accurately distinguish between false and real news tweets. The analysis of the model begins with a confusion matrix - an essential tool to evaluate the performance of classification algorithms. This matrix comprises four essential values:

True Positives (T P), False Positives (F P), True Negatives (T N), and False Negatives (F N), where a positive prediction indicates a fake news tweet. These measures act as the base for well-known performance metrics:

Accuracy (A) - the proportion of correctly classified instances relative to all instances in the dataset. It is calculated using the formula:
A = TN + TPTN + TP + FN + FP
It provides an overall measure for the model's correctness in its predictions.

Precision - The proportion of correctly classified positive predictions (in this case fake tweets) relative to the total positive predictions.
Precision = TPTP + FP
It shows the model’s ability to accurately classify fake tweets without mislabelling real tweets.

Recall - The proportion of correctly classified positive predictions relative to all the actual positive instances.
recall = TPTP + FN
It reflects the model’s ability to identify all the fake news tweets present in the dataset.

F1 Score - The F1 score represents a harmonic mean between precision and recall, offering a balanced assessment of the model’s performance.
F1 score = 2 (Precision Recall)Precision + Recall
Since it considers both false positives and false negatives it proves to be a valuable metric to obtain an overarching view of the model’s performance.

These performance metrics, derived from the confusion matrix, allow us to assess the XGBoost model’s ability to distinguish between real and fake news tweets effectively. Furthermore, they enable comparisons with other state-of-the-art fake news detection methods and provide insights into areas for potential improvement. Results are presented in tables and graphs, complemented by statistical analysis to determine the significance of observed differences.

Results:-

The proposed model demonstrates proficiency in classifying previously unseen tweets, distinguishing them as real or fake news. The model was evaluated against a dataset comprising 13420 previously unseen tweets from the CIC Dataset. A confusion matrix (Figure 7) was used to calculate the metrics that are outlined in 3.1.

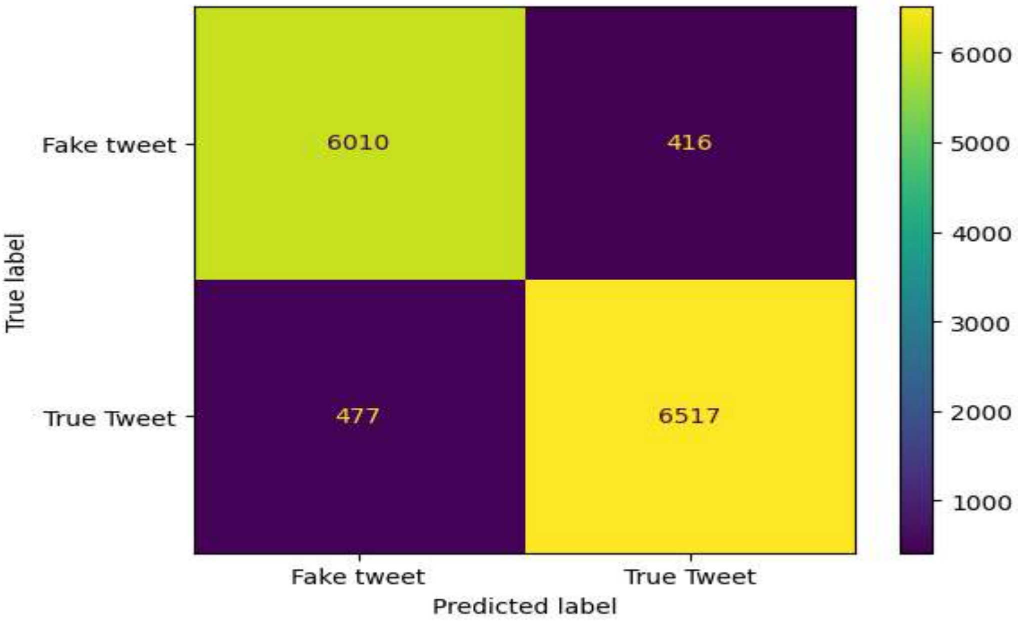

Figure 7:- Confusion Matrix for Results of Proposed Model.

The results outlined in Table 3 underscore the model's capability to effectively classify tweets, distinguishing between tweets with fake and real news with remarkable precision. Notably, the recall value (0.9353) emphasizes the model's proficiency in correctly identifying a substantial proportion of actual fake tweets. The model's accuracy of .9883, while maintaining a high True Positive Rate (TPR) and low False Positive Rate (FPR) (of about 0.1), further highlights its robustness in distinguishing between fake and real news tweets. A high f1 score of 0.8390 indicates the model's equilibrium between precision and recall.

Table 3:- Performance Metrics.

Metric	**Value**
Accuracy	0.9335
Precision	0.9265
Recall	0.9353
F1 Score	0.9360
Area Under Precision-Recall curve	0.953
Area Under an ROC	0.957

Along with the confusion matrix other metrics such as precision- recall curves, and ROCs may also be utilized to evaluate the performance of the model.

The precision-recall curve (Figure 8) shows the trade-off between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

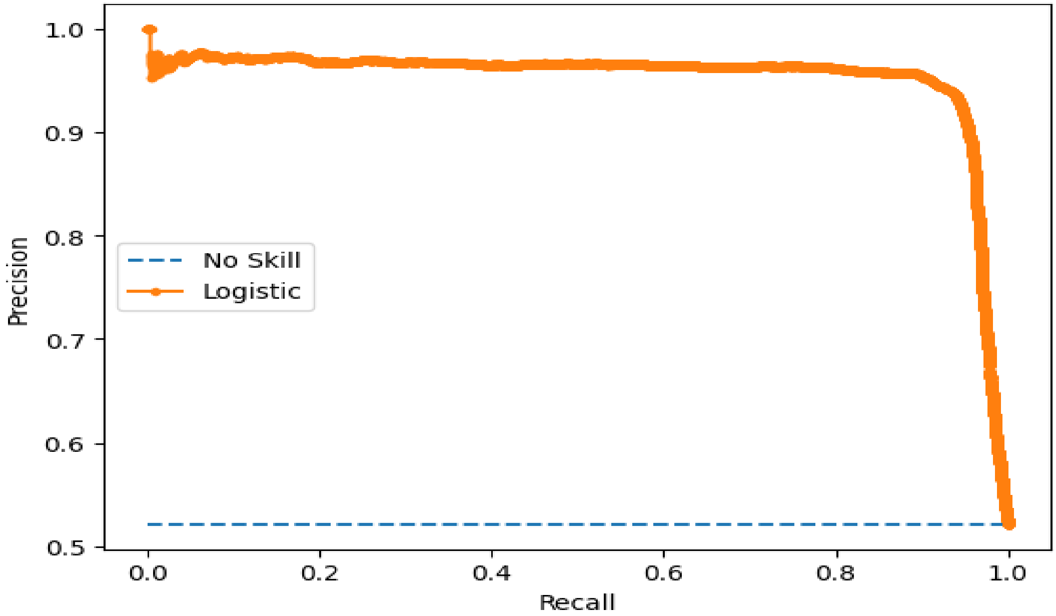

Figure 8:- Precision-Recall curve for the XGBoost model, showing the trade-off between precision and recall. The high area under the curve demonstrates the model's ability to return accurate and relevant results.

A high area under the curve of 0.953 for our model shows that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

A Receiver Operator Characteristic (ROC) curve is a graphical plot used to show the diagnostic ability of binary classifiers. The ROC curve shows the trade-off between sensitivity (TPR) and specificity (1 – FPR). Classifiers that give curves closer to the top-left corner indicate a better performance. A commonly chosen metric to evaluate ROCs is the area under the curve which is equivalent to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance.

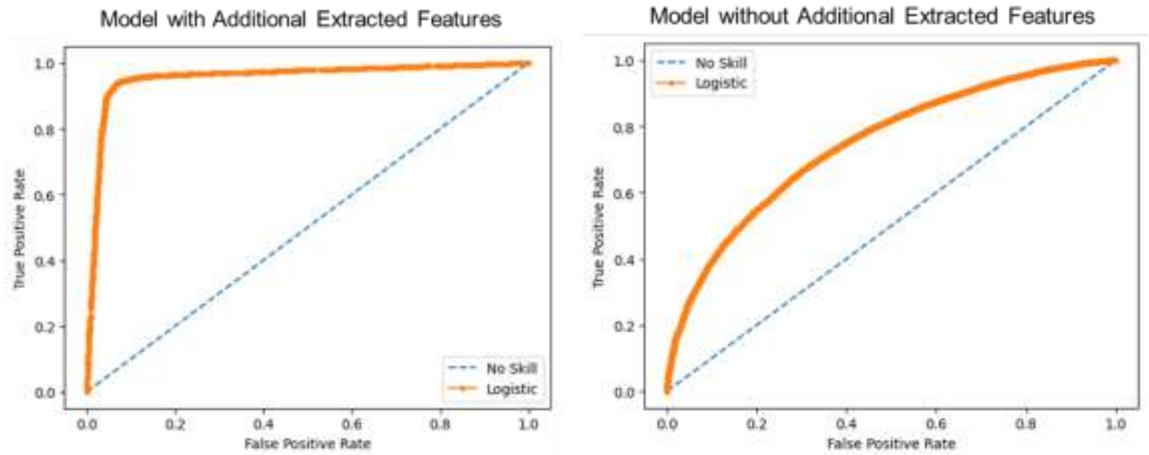

Figure 9:- Comparison of ROC curves for the XGBoost model with additional extracted features and the model without any additional features, showing the trade-off between sensitivity and specificity. The high area under the curve for our model demonstrates the efficacy of the additional extracted features.

One of the unique features of this model is the extraction of additional features from the existing features, as elucidated in the dataset section.

In Figure 9 ,The ROC Curve for our model is compared with the ROC curve for a similar XGBoost model without our additional extracted features, described in the dataset section. The area under the curve of 0.957 for our model is high, indicating greater predictive accuracy when compared to the other model having an AUC of 0.748. Thus, this indicates the high accuracy of our model as well as the efficacy of the additional extracted features such as sentiment scores of the tweets and statements and the features relating to the authors, which significantly contribute to improving the performance of our model.

Collectively, these metrics illustrate the model's potential for practical deployment in tweet detection, particularly concerning the identification of fake news, where timely detection and mitigation are paramount.

Discussion:-

The successful implementation and evaluation of my XGBoost model in the task of classifying fake news and real news on twitter exemplifies the promising potential of machine learning techniques in the domain of fake news detection on social media. In this section, I compare the model against other methods and models that employ similar techniques, discuss the strengths and limitations of this approach, and suggest potential areas for further investigation.

Comparison with Similar Methods

This subsection draws comparisons with the studies that utilize similar approaches to address the proposed problem.

In comparison to Ahmad et al.'s [26] work, which achieved an accuracy of 0.89 by using an XGBoost model on a combined dataset of fake and real news articles ,comprising two datasets from Kaggle and one dataset extracted from the World Wide Web, the proposed XGBoost model significantly outperforms it across all categories, showcasing its distinguished performance for the task of fake news detection. While efficient on individual datasets, their model demonstrated less accuracy when tested on a combined dataset with more variety of tweets and volume such as the CIC Truthseeker 2023 dataset used to train and test the XGBoost model in this research.

Ajao et al.'s work [27], also concerns text sentiment analysis for fake news detection in twitter posts and utilizes an XGBoost model to produce an accuracy of 0.84. My proposed model contributes to this existing research landscape by expanding upon the application of sentiment analysis in fake news detection. While Oluwaseun Ajao et al.'s work [27] only computes the negative and positive polarity scores for the tweets, my model also computes the sentiments for the statements, and the difference between the sentiments for the tweets and the statement, providing an improved and high-performing solution for fake news detection, demonstrating its potential applicability in preventing the spread of misinformation online.

Strengths and limitations

Strengths:

- **Additional Feature Extraction**: The preprocessing of data in the proposed model includes extracting new additional features, such as in-depth sentiment analysis for both the tweet and the statement, along with features related to the author of the statement. These diverse and robust features significantly enhance the XGBoost model's accuracy in classifying and detecting fake news tweets, as demonstrated in Figure 9.
- **Automated Hyperparameter Tuning**: This method involves automated hyperparameter tuning using a grid search algorithm to optimize the model for the dataset's specific characteristics. This flexibility ensures the model achieves optimal performance.
- **Validation-Test Split**: By harnessing a validation-test split, our model can adapt to different features within the tweets, rendering it versatile and adaptable to various types of tweets online. Moreover, by separating the validation and test data the model allows us to evaluate its performance on unseen data, and reduces the overfitting of training data, which is a common drawback of using XGBoost models.

Limitations:

- **Limited Real-World Evaluation:** Although the model demonstrates promising results in controlled testing environments, it has yet to be evaluated in real-world settings. The dynamics of fake news propagation in such contexts can be more unpredictable and influenced by various external factors, potentially impacting the model's effectiveness.
- **Dataset Dependency**: The performance of this model heavily depends on the quality and variety of the training dataset. The model's complexity and the extensive feature extraction process increase the risk of overfitting, making it less generalizable to new, unseen data, enhancing the robustness of the model necessitates including more diverse data, especially since XGBoost models tend to overfit to the specific features of the training dataset. Additionally, the model is limited to detecting tweets in the English language and cannot be applied to multiple languages, as it is specifically tailored to the features of fake news in English.
- **Feature dependency** - The model's performance relies heavily on the availability of supplementary information about the related news article and its author. If such additional information is unavailable, the model may not perform to its demonstrated potential.

Future Work

Though this study has high promises and outcomes, there are still critical considerations regarding the impact of data preprocessing techniques and other decisions chosen in this model.

While my current methodology, which includes extensive feature extraction and hyperparameter tuning, has yielded positive results, there remains room for further exploration. Evaluating alternative preprocessing techniques, such as data augmentation, and dimensionality reduction techniques like Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbour Embedding (t-SNE), could enhance model performance. The extensive hyperparameter tuning and feature extraction used in this study may lead to overfitting to a single dataset. To address this, testing on more diverse datasets or using larger datasets aggregated from various sources on the web would help make the model more generalizable. Additionally, experimenting with alternative machine learning models, or incorporating advanced deep learning techniques like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), could further improve the model. Moreover, adding new features, such as timestamp data and network analysis features such as user interaction graphs, could increase the model's robustness and provide insights into how tweet propagation over time affects detection and distinction characteristics.

Furthermore, although the model exhibits proficiency within a controlled laboratory environment and a structured dataset, there is ample scope to test it in the real world, where the complexities of fake news are continually evolving. This real-world testing will help further refine the model. Additionally, adapting the model for real-time data streams can enhance its applicability as a potential detection system to detect and block potentially harmful fake news. While pursuing these advancements, it is crucial to consider the ethical implications and potential misuse of such detection systems, which could lead to censorship or unintended biases in content moderation practices.

Finally, Addressing the inherent challenge of emerging deception strategies and increasing sophistication, characterized by novel and previously unseen patterns, is imperative for ongoing research. While machine learning models excel under training and evaluation conditions that mirror known patterns, the dynamic nature of fake news detection necessitates regular model updates to effectively accommodate emerging threats.

Conclusion:-

This study underscores the potential of advanced machine learning techniques, specifically gradient-boosted decision trees, in the realm of fake news detection on social media platforms. By leveraging the TruthSeeker 2023 dataset and incorporating features such as sentiment analysis and author characteristics, our model achieved a high accuracy of 0.9335 on unseen tweets, surpassing many existing methodologies in key metrics.

Our approach's adaptability, versatility, and generalization capabilities position it as a promising candidate for real-world deployment, where the timely identification and mitigation of misinformation are crucial. Fake news poses a significant challenge not only to social media platforms but to the broader information ecosystem as well.

In this work, we have presented a robust detection architecture that not only offers an effective solution for fake news detection but also advances the field of misinformation research in the digital age. The strong performance of our model paves the way for enhanced information integrity, protecting users and platforms from the harmful effects of fake news. As the landscape of misinformation continues to evolve, ongoing research and ethical deployment of such technologies will be essential in safeguarding the truth and fostering a more informed society.

References:-

[1] Velichety, S., & Shrivastava, U. (2022). Quantifying the impacts of online fake news on the equity value of social media platforms – Evidence from Twitter. International Journal of Information Management, 64, 102474. https://doi.org/10.1016/j.ijinfomgt.2022.102474

[2] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. Science, 363(6425), 374–378. https://doi.org/10.1126/science.aau2706

[3] Howell, W. L. (Ed.). (2013). Global Risks 2013. World Economic Forum. https://www3.weforum.org/docs/WEF_GlobalRisks_Report_2013.pdf

[4] Spencer, D. R. (2007). The Yellow Journalism: The Press and America's Emergence as a World Power. United Kingdom: Northwestern University Press. https://www.google.co.in/books/edition/The_Yellow_Journalism/d79fyolBDgAC?hl=en

[5] Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2021). An exploratory study of COVID-19 misinformation on Twitter. Online Social Networks and Media, 22, 100104. https://doi.org/10.1016/j.osnem.2020.100104

[6] Al-Rawi, A., O'Keefe, D., Kane, O., & Bizimana, A. J. (2021). Twitter's Fake News Discourses Around Climate Change and Global Warming. Frontiers in Communication, 6. https://doi.org/10.3389/fcomm.2021.729818

[7] Kumar, S., & Shah, N. (2018b, April 23). False Information on Web and Social Media: A Survey. arXiv.org. https://arxiv.org/abs/1804.08559

[8] Yamak, Z., Saunier, J., & Vercouter, L. (2018). SocksCatch: Automatic detection and grouping of sockpuppets in social media. Knowledge-based Systems, 149, 124–142. https://doi.org/10.1016/j.knosys.2018.03.002

[9] Bello, B. S., Heckel, R., & Minku, L. (2018). Reverse Engineering the Behaviour of Twitter Bots. https://doi.org/10.1109/snams.2018.8554675

[10] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

[11] Bodaghi, A., & Oliveira, J. (2022). The theater of fake news spreading, who plays which role? A study on real graphs of spreading on Twitter. Expert Systems With Applications, 189, 116110. https://doi.org/10.1016/j.eswa.2021.116110

[12] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017, August 23). Automatic Detection of Fake News. arXiv.org. https://doi.org/10.48550/arXiv.1708.07104

[13] Choi, D., Chun, S., Oh, H., Han, J., & Kwon, T. T. (2020). Rumor Propagation is Amplified by Echo Chambers in Social Media. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-019-57272-3

[14] CAROL SOON, SHAWN GOH (2018-09). FAKE NEWS, FALSE INFORMATION AND MORE: COUNTERING HUMAN BIASES : 1-71. ScholarBank@NUS Repository. https://doi.org/10.25818/t406-zy24

[15] Graham, C., & Blankenship, M. (2020, July 6). How misinformation spreads on Twitter. Brookings. https://www.brookings.edu/articles/how-misinformation-spreads-on-twitter/

[16] Misinformation and disinformation. (2023, November 29). https://www.apa.org. https://www.apa.org/topics/journalism-facts/misinformation-disinformation

[17] T, S. M., & Mathew, S. K. (2022). The disaster of misinformation: a review of research in social media. International Journal of Data Science and Analytics, 13(4), 271–285. https://doi.org/10.1007/s41060-022-00311-6

[18] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying Misinformation in Microblogs. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1589–1599, Edinburgh, Scotland, UK.. Association for Computational Linguistics. https://aclanthology.org/D11-1147

[19] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013. https://doi.org/10.1145/2487788.2488033

[20] Horne, B. D., & Adali, S. (2017, March 28). This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. arXiv.org. https://doi.org/10.48550/arXiv.1703.09398

[21] Acemoglu, D., Ozdaglar, A., & ParandehGheibi, A. (2010). Spread of (mis)information in social networks. Games and Economic Behavior, 70(2), 194–227. https://doi.org/10.1016/j.geb.2010.01.005

[22] Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). Epidemiological modeling of news and rumors on Twitter. https://doi.org/10.1145/2501025.2501027

[23] Tripathy, R. M., Bagchi, A., & Mehta, S. (2010). A study of rumor control strategies on social networks. https://doi.org/10.1145/1871437.1871737

[24] Truth Seeker Dataset 2023 | Datasets | Research | Canadian Institute for Cybersecurity | UNB. (n.d.). https://www.unb.ca/cic/datasets/truthseeker-2023.html

[25] Figure 2. Simplified structure of XGBoost. (n.d.). ResearchGate. https://www.researchgate.net/figure/Simplified-structure-of-XGBoost_fig2_348025909

[26] Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake News Detection Using Machine Learning Ensemble Methods. Complexity, 2020, 1–11. https://doi.org/10.1155/2020/8885861

[27] Ajao, O., Bhowmik, D., & Zargari, S. (2019). Sentiment Aware Fake News Detection on Online Social Networks. https://doi.org/10.1109/icassp.2019.8683170

Appendix:-

The dataset used in this study consists of 130,000 tweets collected from Twitter. These tweets were labelled as either fake news or real news based on the criteria described in 2.1. Table A1 provides a summary of the features extracted from each tweet.

Table A1:- Detailed Dataset Description.

Feature	Description
Tweet Features	
statement	Headline of the news article
author	The author of the statement
target	The ground truth value of the statement
BinaryNumTarget	Binary representation of the target value (1 = True, 0 = False)
Manual keywords	manually created keywords used to search twitter with
Tweet	twitter posts related to the associated manual keywords
Text features	
Unique count	Number of Unique, complex words in the tweet
Total Count	Total number of words in the tweet
Max Word Length	Length of longest word in the tweet
Average Word Length	Average length of words in the tweet
Lexical Features	
Present verbs	Number of present tense verbs in the tweet
adjectives	Number of adjectives in the tweet
capitals	Number of Capitalized letters in the tweet
exclamations	Number of (!) used
questions	Number of (?) used
Long word frequency	Number of long words

Metadata features	
Followers count	Number of followers of user
Friends count	Number of friends of user
replies	Number of replies the user has
retweets	Number of retweets the user has
favourites	Number of favourites the user has
hashtags	Number of hashtags the user has used
BotScoreBinary	Binary score whether the user is considered a bot or not
cred	Credibility score
Additional Extracted features	
Number of authors	Total number of unique authors
Fraction of long words per tweet	Fraction of number of long words per total number of words in a tweet
Sentiment of the tweet	Compound sentiment score of the tweet
Sentiment of the statement	Compound sentiment score of the statement
Difference between tweet and statement sentiment	Difference between compound sentiment score of the statement and compound sentiment score of the tweet
Average sentiment per author	Average compound sentiment for each author
Average statement length per author	Average statement length for each author