Journal of Theoretical and Applied Information Technology
15th January 2024. Vol.102. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCING FAKE NEWS DETECTION ON SOCIAL MEDIA THROUGH ADVANCED MACHINE LEARNING AND USER PROFILE ANALYSIS

ZAHIR ABBAS KHAN[1], REKHA V[2]

[1] Research Scholar, Department of Computer Science and Engineering, School of Engineering and Technology, Christ University, Bangalore, India

[2] Assistant Professor, Department of Computer Science and Engineering, School of Engineering and Technology, Christ University, Bangalore, India

E-mail:[1]zahir.khan@res.christuniversity.in, [2]rekha.v@christuniversity.in

ABSTRACT

Social media news consumption is growing in popularity. Users find social media appealing because it's inexpensive, easy to use, and information spreads quickly. Social media does, however, also contribute to the spread of false information. The detection of fake news has gained more attention due to the negative effects it has on society. However, since fake news is created to seem like real news, the detection performance when relying solely on news contents is typically unsatisfactory. Therefore, a thorough understanding of the connection between fake news and social media user profiles is required. In order to detect fake news, this research paper investigates the use of machine learning techniques, covering important topics like feature integration, user profiles, and dataset analysis. To generate extensive feature sets, the study integrates User Profile Features (UPF), Linguistic Inquiry and Word Count (LIWC) features, and Rhetorical Structure Theory (RST) features. Principal Component Analysis (PCA) is used to reduce dimensionality and lessen the difficulties presented by high-dimensional datasets. The study entails a comprehensive assessment of multiple machine learning models using datasets from "Politifact" and "Gossipofact," which cover a range of data processing methods. The evaluation of the XGBoost classification model is further enhanced by the analysis of Receiver Operating Characteristic (ROC) curves. The results demonstrate the effectiveness of particular combinations of features and models, with XGBoost outperforming other models on the suggested unified feature set (ALL).

Keywords: *Linguistic Inquiry and Word Count features, Machine learning techniques, Principal Component Analysis (PCA), Rhetorical Structure Theory (RST) features, Unified feature set (ALL), User profile features (UPF)*

1. INTRODUCTION

Disinformation, often colloquially referred to as "fake news," is a phenomenon that has been present throughout human history, although the term and its current manifestations are relatively new. While the term "fake news" gained widespread recognition in the 21st century, the concept of spreading false or misleading information predates the internet and modern media by centuries. Here, we'll explore the historical roots of disinformation, its evolution, and its contemporary impact[1]. As evidenced by the 2016 U.S. presidential election, where Macedonian youths operated over 100 websites disseminating false information, fake news, which is frequently misinformation, is frequently intentionally created for political, personal, or financial gain. Fact-checking is done by the media and experts to thwart false information. Manual fact-checking, however, is a labor and expertise-intensive task that cannot be scaled to the volume of fake news generated. As a result, researchers have developed data mining and machine learning techniques for automatically spotting fake news [2]. Some examples of fake news which affect the environment, in 2017, reports surfaced that Barack Obama had suffered injuries from an explosion [3]. On occasion, the dissemination of false information has as its goal escalating the anxiety and causing confusion, as was the case with the "Pizzagate," which prompted a man to

assault a restaurant in 2016 with a rifle false information that in the restaurant, there were young children as well as sex slaves in a child abusing led were widely published by Hillary Clinton across the web $130 billion in stock value was lost [4]. In light of this, we can state that the scope, variety, and serious risks of fake news, as well as the more general threat that the misinformation being spread on social media becoming cause for worry as a result of the Potential social cost that it could have soon future[5]. Due to all of these factors, identifying fake news spreaders and detecting fake news in general have become crucial tasks, the primary goal of this proposal is to analyze what function user-profiles and other features play in such a task.

Recent studies have shown that examining the relationship between user profiles and the dissemination of false information can help identify users who are more likely to believe false information and distinguish them from users who are more likely to believe true information [6]. Due to the encouraging outcomes found in recent studies, approaches that take context-based features in addition to content-based features have been obtaining prominence in recent years [7].

In a time when false information is abundant, understanding the origins and development of disinformation is essential to understanding its enduring nature and shifting forms. The deliberate dissemination of false information during crucial occasions, like the 2016 U.S. presidential election, highlights the significant influence of fake news on public opinion and significant societal advancements. Problems associated with manual fact-checking demonstrate how unworkable conventional methods are, leading to a move towards data-driven solutions. The tangible risks posed by misinformation are highlighted by incidents of false reports impacting different facets of society and inciting actual violence, underscoring the necessity of proactive detection and mitigation techniques. The extensive range, hazards, and possible societal repercussions linked to false information underscore the pressing need to put policies in place to combat disinformation and safeguard public welfare.

The widespread problem of misinformation, sometimes referred to as "fake news," is the issue or knowledge gap that this study attempts to fill. The phenomenon of disseminating inaccurate or misleading information has historical roots that predate the internet and contemporary media, even though the term only became well-known in the twenty-first century. Disinformation's development presents a serious problem, especially when considering its current influence. The study's importance stems from the growing influence of fake news, which has been demonstrated by events like the 2016 U.S. presidential election and cases where inaccurate information resulted in negative outcomes such as financial losses and violent crimes. Given the volume of fake news produced, the media's traditional manual fact-checking procedure is both labor-intensive and impractical. Thus, it is imperative to investigate and create automated strategies to identify and stop the spread of false information, such as data mining and machine learning approaches.

Studying this type of information primarily entails three aspects:

- Information about the user, including age, place of residence, the number of followers, etc.
- Since users often use comments to share their views and since they can aid in the negative aspects, responses generated by fake news can serve as a vital means of detection of a user credibility index [8].
- The social media platforms used to spread news because the quick propagation of those networks is employed to connect with the greatest range of users in a brief period of time, the study of the networks through which information is propagated has particular relevance.

The specific contributions of this research involve the analysis of the role played by user profiles and other features in identifying fake news spreaders and detecting fake news in general.

We made several important improvements in the fields of data analysis and machine learning in this research paper. The following can be used to summaries these contributions:

a) Feature Engineering and Selection: We combined different feature categories, such as Linguistic Inquiry and Word Count features (LIWC), User Profile features (UPF), and Rhetorical Structure Theory (RST) features, in an exacting process of feature engineering and selection. We were able to get a more complete picture of the

data thanks to the combination of these features.

b) Dimensionality Reduction: Following feature selection, we used Principal Component Analysis (PCA) to address the problem of high dimensionality in our datasets. This dimensionality reduction method reduced the likelihood of overfitting while simultaneously improving computational efficiency.

c) Extensive Model Assessment: Using the "Politifact" and "Gossipofact" datasets, we carried out an extensive assessment of a number of machine learning models, including Bagging, Random Forest, Extra Tree, AdaBoost, Gradient Boosting, and XGboost. Several feature sets and data processing techniques (OD, WRST, and PCA) were used in this evaluation.

d) ROC Curve Analysis: We used Receiver Operating Characteristic (ROC) curves on three different datasets to evaluate the performance of the XGBoost classification model. These curves give a graphical depiction of the model's capacity to classify positive and negative examples in a variety of scenarios.

e) Model and Feature Selection: We found the best model feature combinations based on the thorough evaluation; XGboost performed the best in terms of accuracy on the suggested Unified feature set (ALL).

2. LITERATUREREVIEW

A profiling-avoiding algorithm is proposed that uses Twitter users during model optimization to exclude them from evaluating an article's veracity. The algorithm uses objective functions to maximize correlation between the article and its spreaders. The algorithm has been applied to three neural classifiers, showing positive performance on fake news data and better discrimination between fake and true news [9]. An interactive research paper recommender system uses a topic model to analyze themes in a researcher's work and learn user preferences based on feedback. The system models query as a bag of topics, estimating similarity between queries and papers using a bag-of-topics measure. Users' preferences are considered by tracking favorite papers, incorporating recurring topics, and truncating ignored topics [10]. The paper proposes a method for detecting fake news on social networking sites using biased words in self-descriptions. Using machine learning, feature vectors are created from multiple users posting the same news URL. The method achieved an average classification accuracy of 90.2% in experiments using real and fake news from Japan and the U.S[11]. Social media has enabled the dissemination of fake news, causing confusion and disruptions on society. To combat this, computational fake news detection research has gained attention. However, the lack of comprehensive and community-driven fake news benchmark datasets is a major hurdle. This paper presents FakeNewsNet, a fake news benchmark data repository with two comprehensive datasets with diverse features in news content, social context, and spatiotemporal information [12]. This article reviews the state-of-the-art approach to user profiling, focusing on methods, characteristics, and taxonomy of user profiles. It discusses data acquisition, feature extraction, profiling techniques, and performance measures. Challenges include privacy, datasets, cold start issues, trust issues, and computational complexity. The article also highlights an open research direction for further research in advancing user profiling. The findings show that an effective modeling process enhances the construction of accurate user profiles for service personalization [13].

This study investigates review-based user profiling (RBUP), a new research direction, by mapping 51 out of 2478 papers. The researchers identified a generic process for RBUP and performed multi-dimensional analysis on each step. The results show that traditional methods are not sufficient to fully understand user characteristics and requirements, and that further investigation is needed. The study highlights the need for systematic review and further research [14]. The World Wide Web offers millions of users access to information, entertainment, and e-commerce. However, it's difficult to categorize users based on their preferences. A solution is to create a web-platform that acts as a middleware, analyzing user data using 'user profiling'. This article presents an online profiling mechanism in a virtual e-shop, demonstrating how neural networks can predict new user characteristics. This approach benefits both customers and stores by delivering

personalized advertisements directly to mobile devices [15]. A new model predicts user influence in spreading fake or real news on social media, considering user and tweet characteristics. It outperforms existing models and investigates important features for predicting real versus fake news spreaders [16]. Fake news has become a significant issue on social media due to its rapid spread and potential to change public opinion. This article analyzes two Twitter datasets to identify users who share fake news. The study reveals that user characteristics, such as personality traits, emotions, and writing style, are strong predictors of fake news spreaders, outperforming baseline models and achieving an average precision of 0.80 to 0.99 [17]. The study introduced a multiple imputation strategy using Multiple Imputation Chain Equation (MICE) to handle missing variables in social media or news data. Term Frequency and Inverse Document Frequency were used to extract effective features. Naïve Bayes, passive-aggressive, and Deep Neural Network (DNN) classifiers were used to classify missing data variables. The method achieved 99.8% accuracy in detecting fake news [18]. The study used sentiment and emotion analysis of news articles and user comments to develop a bidirectional long short-term memory model for detecting fake news. The model achieved a high detection accuracy of 96.77%, surpassing other studies. The extracted features, based on publisher and crowd stances, improved the model's efficiency [19]. SceneFND is a system that combines textual, contextual, and visual representations to detect fake news. It uses word embeddings and images from news posts. Statistical analysis reveals significant differences in fake and real news frequency. Experimental results show SceneFND improves textual baseline performance by 3.48% in PolitiFact and 3.32% in GossipCop datasets.

Numerous research gaps or questions in the fields of fake news detection, user profiling, and related fields can be found based on the information provided in the text. Here are some possible research questions and gaps:

a) Feature Integration for the Identification of False News: To create a comprehensive and informative feature set that will improve the performance of fake news detection systems, there is a lack of established methodologies and best practices for effectively integrating different feature categories, such as linguistic, user profile, and structural features.

b) Handling High Dimensionality in Datasets for Detecting False News, Research Gap: The methods that can be utilized to overcome the problem of high dimensionality in datasets used for fake news detection are not well understood by the research community at this time. Further research is necessary to determine the effects of these methods on computational effectiveness and overfitting risk in these kinds of systems.

c) Entire Model Assessment for False News Identification: We don't fully understand the important parameters to take into account when assessing machine learning models for the detection of fake news in a variety of feature sets and datasets. To find out how these elements can influence the choice of ideal models and configurations for trustworthy detection systems, more research is required.

d) Model Performance Assessment Using Visual Analysis: To improve the comprehension of machine learning model performance in the context of fake news detection, there is a dearth of thorough investigation in the field regarding the application of visual analysis techniques, such as ROC curve analysis. Comprehending the significance and possible benefits of visual evaluations is still an uncharted territory.

e) Ideal Model-Feature Combinations for the Identification of False News: At the moment, little research has been done to pinpoint and clarify the feature sets and model combinations that work best together to achieve high levels of interpretability and accuracy in the detection of fake news. Moreover, more research is needed to determine the best practices for finding and using these combinations to improve system performance.

3. PROPOSED METHODOLOGY

The deliberate goal of the research is to disentangle the complex relationship that

Journal of Theoretical and Applied Information Technology
15th January 2024. Vol.102. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

exists between social media user profiles and fake news. The present investigation utilizes machine learning methodologies, encompassing essential aspects like feature integration, user profiles, and dataset analysis.

The methodological approach is broad, building large feature sets with features from Rhetorical Structure Theory (RST), Linguistic Inquiry and Word Count (LIWC), and User Profile Features (UPF). Principal Component Analysis (PCA) is also used to lessen the difficulties posed by high-dimensional datasets.

Figure 1 shows the suggested methodology. First, the basis of this research is built on the FakeNewsNet dataset, which offers a wide range of labeled real and fake news articles and social media posts. Pre-processing techniques are used to get the data ready for analysis, like text cleaning, tokenization, and lemmatization.

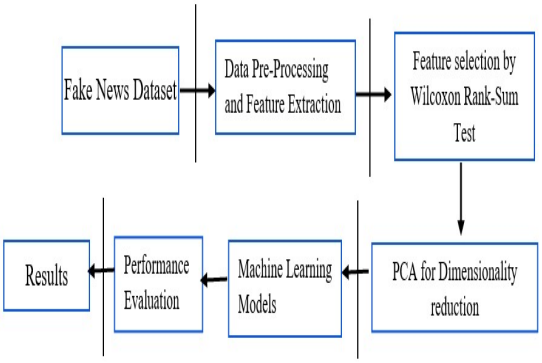

Figure 1: Proposed Architecture diagram of user Profile Features.

3.1 Dataset Description

The research utilized the FakeNewsNet dataset, which includes data from PolitiFact and Gossipcop platforms, to identify which user sharing fake news, table 1 shows dataset includes user count, shared articles, and articles classified as true and fake news.

Table I: FakeNewsNet Dataset Details

Dataset	Platform	Users (without filtering bots)	Sharing	True News	Fake News
FakeNews Net	Politifact	159,699	271,462	361	361
FakeNews Net	Gossipcop	209,930	812,194	4,513	4,513

3.2 Data Pre-processing and feature extraction

We enumerate the critical steps in feature extraction and data preprocessing. Figure 2 display the procedure of creating a new feature with the help of feature extraction process.

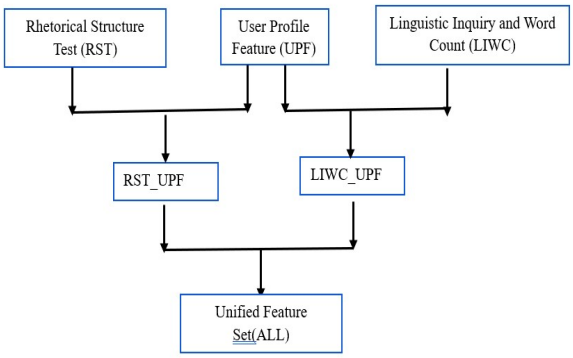

Figure 2. Feature extraction process.

3.2.1 User Profile features

User profile features are an essential part of our research because they shed light on the traits of users who interact with news articles. These characteristics are taken from user profiles linked to news sources and content. The goal of our research and the significance of user profile features. Features of user profiles assist in identifying user preferences and behavior patterns, which may indicate whether a news source or article is trustworthy or authentic. For example, specific user behaviors, like posting trends and levels of engagement, may point to the existence of fake news or biased reporting. Incorporating user profile features will help our classification model be more discriminatory. These features give the model a more thorough understanding of the context surrounding news articles, allowing it to make better decisions.

3.2.2 Rhetorical structure theory (RST) features

Rhetorical Structure theory is widely used in fields like linguistics, discourse analysis, and natural language processing (NLP) to understand and generate coherent text. It is particularly useful for tasks such as text

summarization, text generation, and discourse analysis because it provides a framework for modeling the relationships and structure within a given text[21].They assist in determining how information is organized within articles, including headings, subheadings, and their relationships, which can be a sign of good journalism[22].The ability of our model to distinguish between articles with organized, trustworthy content and those with disorganized or misleading content is improved by the inclusion of RST features.

3.2.3 LIWC Features

Features like Linguistic Inquiry and Word Count (LIWC)[23] offer linguistic and psychological insights into news article text. These characteristics, which come from linguistic analysis, are important for comprehending the tone, sentiment, and writing style of news articles. We can extract linguistic and emotional cues from news articles thanks to LIWC features. This can aid in identifying potential biases, polarities of sentiment, and linguistic traits that might be connected to false information or unreliable reporting. The ability of our model to analyze the textual content of news articles is enhanced by incorporating LIWC features, enabling it to capture nuances beyond straightforward keyword analysis.

RST_UPF [24]: The concatenated features of RST and UPF, which comprise features retrieved from user profiles and news content alike.

LIWC_UPF [24]: The combined features of LIWC and UPF, which incorporate elements taken from user profiles and news content, are represented by LIWC_UPF.

UFS (ALL): The "UFS" feature set includes every feature that is accessible, offering a thorough and multi-dimensional analysis of the news content of dataset.

In our research, the Unified Feature Set (UFS) represented a holistic approach to fake news detection and analysis. By integrating these diverse features, we aimed to improve the accuracy and robustness of our models, with the goal of reducing the likelihood of false positives and false negatives. This comprehensive analysis allowed us to adapt to evolving fake news strategies and enhance our understanding of this critical issue. UFS played a pivotal role as an innovation in our research, contributing to more effective fake news detection and classification.

$$Unified\ Feature\ set = \cup \{F(i): i \in UPF, RST, LIWC, (RST + UPF), (LIWC + UPF), ALL\} \quad (1)$$

Where,

F(i) represents the individual feature set.

i∈ {UPF, RST, LIWC, (RST + UPF), (LIWC + UPF), ALL}: This part specifies the set of feature sets being integrated into the UFS. The UFS encompasses feature sets like User Profile Feature (UPF), Rhetorical Feature (RST), Linguistic Inquiry and Word Count (LIWC), and combinations of these sets, as well as the all-inclusive "ALL" feature set.

3.3.4 Feature selections

A feature selection technique based on p-values is employed prior to Principal Component Analysis (PCA). This technique systematically assesses the significance of each feature in the dataset by calculating p-values using the Wilcoxon rank-sum test, which determines whether there is a substantial distinction in feature values between different classes.

The following steps for feature selection before PCA:

Algorithm 1: Feature selection algorithms

Input:

- Dataset**:** Loaded from a CSV file, containing features and target labels.
- Significance Threshold (alpha**):** A predefined threshold (e.g., 0.05) for determining feature significance.

Output:

- Relevant Feature Indices**:** A list of indices corresponding to features deemed significant based on p-values (**alpha**).
- Reduced Dataset**:** A dataset containing only relevant features and target labels, used for PCA and modeling.
- Saved Dataset**:** The reduced dataset is saved for further analysis and modeling.

Procedure:

1. Load Data:
   - Load the dataset from a CSV file for the selected dataset ("politifact" or "gossipcop").
2. Initialize Variables:
   - Set a significance threshold alpha (e.g., 0.05).
   - Create an empty list *relevant features* to store relevant feature indices.
3. For Each Feature in the Dataset:

- Iterate through each feature *X[i]* in the dataset.

4. Calculate P-value:
   - Calculate the p-value *p_value* for *X[i]* using the Wilcoxon rnk-sum test.
5. Check P-value Significance:
   - If *p_value*<= alpha:
   - The feature is considered relevant.
   - Add *i* to *relevant_features.*
6. Create Reduced Dataset:
   - Extract the relevant feature *X[relevant_features]* and target labels *y* from the dataset.
7. Save Reduced Dataset:
   - Save the reduced dataset with relevant features and target labels into new CSV files.

End:

Features with p-values below a specified significance threshold (e.g., 0.05) are considered statistically relevant and are retained, while those with higher p-values are deemed irrelevant and are excluded from further analysis. The objective of this feature selection step is to reduce the dataset's dimensionality while preserving only the features that demonstrate statistical significance, ultimately enhancing the efficiency and effectiveness of subsequent PCA dimensionality reduction and machine learning model training processes. The choice of the significance threshold allows for customization based on the desired level of statistical confidence in the selected features.

3.3 Machine learning algorithms for classification

3.3.1 Bagging with K-Nearest Neighbors (KNN)

An ensemble method called bagging aims to increase the accuracy and stability of a base classifier, in this case, K-Nearest Neighbors. By using different instances of the base classifier to make predictions on different subsets of the training data, bagging can make predictions. In order to strengthen KNN's robustness and reduce its propensity for overfitting, Bagging is used in conjunction with KNN. When the distribution of the underlying data is complicated, it can be especially helpful.

$$\hat{y}bagging(x) = \frac{1}{N}\sum_{i=1}^{N} \hat{y}KNN_i(x) \quad (2)$$

Where,

$\hat{y}bagging(x)$: *Prediction bagged for x data point.*
$\hat{y}knn_i(x)$: *the ith KNN model's prediction for data point x.*
N:Number of Knn Models

3.3.2 Random Forest

An ensemble technique called Random Forest constructs numerous decision trees and combines their forecasts. By averaging the outcomes of different trees, it aims to decrease overfitting and enhance generalization. Due to its capability to handle high-dimensional data and provide feature importance rankings, Random Forest is frequently used for classification tasks. Complex relationships between features and labels are well-captured by it.

$$\hat{Y}RF(x) = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_{treei}(x) \quad (3)$$

3.3.3 Extra Trees

A decision tree-based ensemble method, Extra Trees is similar to Random Forest. But how it builds individual trees is different. More randomness is incorporated into the tree-building process by Extra Trees, which can lessen overfitting. When working with noisy or high-dimensional data, Extra Trees can be especially helpful. It is renowned for its sturdiness and effectiveness in handling outliers.

$$\hat{Y}ET(x) = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_{treei}(x) \quad (4)$$

3.3.4 AdaBoost

AdaBoost is an ensemble technique that improves a base classifier's performance by giving samples that were incorrectly classified more weight. In order to concentrate on the most complex cases, it iteratively modifies the weights of the data points. AdaBoost is frequently used with struggling students and can significantly enhance classification performance. It manages class disparity well and can adjust to complex data distribution.

$$\hat{Y}AdaBoost(x) = sign\left(\sum_{i=1}^{N} \alpha_i . \hat{y}_{week_i}(x)\right) \quad (5)$$

3.3.5 Gradient Boosting

Another ensemble method, gradient boosting, constructs decision trees in a sequential manner,

with each tree correcting the flaws of the previous one. Using gradient descent, a loss function is optimized. High predictive accuracy and the capacity to recognize intricate relationships between features and labels are two features of gradient boosting. It is frequently employed in many machine learning applications and contests.

$$\hat{Y}GB(x) = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_{treei}(x \quad (6)$$

3.3.6 GBoost

A highly optimised version of gradient boosting is called XGBoost (Extreme Gradient Boosting). It supports a variety of loss functions and regularization methods and is built for efficiency and predictive performance. Due to its accuracy and speed, XGBoost has gained popularity for classification tasks. It frequently outperforms competing algorithms and is renowned for its prowess in navigating sizable datasets and intricate feature interactions.

$$YXGBoost(x) = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_{treei} \quad (x) \quad (7)$$

3.4 Performance Measurement

The following evaluation metrics are frequently employed in research papers to gauge the effectiveness of classification models:

$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$	(8)
$Roc\ Curve = TPR = \frac{TP}{TP + FN}$	(9)
$FPR\ \frac{FP}{FP + FN}$ $Recall = \frac{TP}{TP + FN}$	(10)
$Specificity = \frac{TN}{(TN + FP)}$	(11)
$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$	(12)

4. EXPERIMENTAL RESULTS

In essence, this research addresses a pressing societal problem – the rampant spread of fake news on social media – by grounding its conceptualization in the imperative for an advanced understanding of user profiles and their role in the dissemination of false information. The results highlight the effectiveness of the proposed approach, with implications for developing more robust and nuanced strategies to counteract the pervasive issue of fake news in the digital era.

X

In this section, we went over the procedures used to export and process the data from our experiment. The preparation of the data for analysis in the future and the production of insights from the information gathered are the main goals. In our experiment, we combined features from fakeNewsDataset for politifact and Gossipofact the LIWC (Linguistic Inquiry and Word Count) features, user profile features (UPF), RST, RST_UPF, LIWC_UPF, UFS (ALL). In the experiment, specific features from the data were chosen for analysis. To concentrate on pertinent elements of the dataset, this was crucial. User features, RST features, and LIWC features were the three feature categories taken into consideration and concatenation has been done with respect to each feature. After this stage starting with the two classes (0 and 1), we applied feature selection techniques based on the Wilcoxon rank-sum test to find and keep features with significant discriminatory power and performed dimensionality reduction by PCA. We first used feature selection to improve the accuracy and relevance of our input data, and then we used Principal Component Analysis (PCA) to lessen dimensionality. The necessity to address high dimensionality issues while preserving important data within the datasets served as the driving force behind this crucial step. A logarithmic calculation based on the initial number of features present in each dataset was used to determine the number of principal components to retain. The result was a notable reduction in dimensionality, which enhanced computational effectiveness and significantly decreased the risk of overfitting.

Table II. Different Machine Learning Models' Performance on Feature Sets and Datasets

Data Set	*Form*	*Features*	*Bagging*	*Random Forest*	*Extra Tree*	*Ada Boost*	*Gradient Boosting*	*XGboost*
Polifact	OD	UPF	0.80758	0.90909	0.89432	0.86212	0.85606	0.90303
Polifact	OD	RST	0.89924	0.8947	0.87689	0.88068	0.88523	0.87273
Polifact	OD	LIWC	0.85173	0.86186	0.8979	0.84497	0.90503	0.90466
Polifact	OD	RST_UPF	0.78679	0.91892	0.91592	0.91892	0.94032	.97698
Polifact	OD	LIWC_UPF	0.75269	0.95507	0.95507	0.90668	0.96966	0.97926
Polifact	OD	UFS (ALL)	0.85119	0.93016	0.92976	0.90754	0.89048	0.978980
Gossipcop	OD	UPF	0.93477	0.99211	0.99197	0.97968	0.97671	0.996760
Gossipcop	OD	RST	0.63209	0.63618	0.6485	0.6011	0.60222	0.62404
Gossipcop	OD	LIWC	0.7765	0.79319	0.78521	0.75683	0.76196	0.84586
Gossipcop	OD	RST_UPF	0.94896	0.99012	0.99171	0.97921	0.9789	0.99648
Gossipcop	OD	LIWC_UPF	0.93088	0.98556	0.983	0.97989	0.98013	0.99479
Gossipcop	OD	UFS (ALL)	0.95002	0.99098	0.98307	0.97682	0.97681	.99791
Polifact	WRST	UPF	0.8948	0.90347	0.93854	0.9144	0.92798	0.94721
Polifact	WRST	RST	0.70155	0.7845	0.76047	0.68488	0.70775	0.77016
Polifact	WRST	LIWC	0.84917	0.86011	0.90234	0.8356	0.84012	0.89593
Polifact	WRST	RST_UPF	0.82424	0.92349	0.90644	0.89659	0.90114	0.94773
Polifact	WRST	LIWC_UPF	0.75901	0.83709	0.86186	0.90015	0.93168	0.93994
Polifact	WRST	UFS (ALL)	0.83411	0.97791	0.97481	0.9876	0.9845	0.98682
Gossipcop	WRST	UPF	0.9136	0.99004	0.99051	0.98148	0.97719	0.9947
Gossipcop	WRST	RST	0.59553	0.6308	0.61211	0.58584	0.58585	0.60569
Gossipcop	WRST	LIWC	0.73328	0.77583	0.74112	0.73921	0.73518	0.80488
Gossipcop	WRST	RST_UPF	0.9363	0.98919	0.99095	0.97743	0.97468	0.99435
Gossipcop	WRST	LIWC_UPF	0.92799	0.98769	0.98212	0.98191	0.98396	0.99677
Gossipcop	WRST	ALL	0.94693	0.99232	0.98983	0.98001	0.98181	0.99694
Polifact	PCA	UPF	0.76486	0.84985	0.82241	0.83842	0.81479	0.80945
Polifact	PCA	RST	0.82331	0.78985	0.8109	0.76391	0.80564	0.8312
Polifact	PCA	LIWC	0.90602	0.88496	0.88421	0.94962	0.89774	0.89925
Polifact	PCA	RST_UPF	0.83797	0.88421	0.87256	0.86203	0.86391	0.87444
Polifact	PCA	LIWC_UPF	0.7783	0.92519	0.86977	0.88411	0.89419	0.91395
Polifact	PCA	UFS (ALL)	0.86967	0.92683	0.93178	0.94474	0.92569	0.93979
Gossipcop	PCA	UPF	0.9252	0.94292	0.94005	0.90723	0.9021	0.94335
Gossipcop	PCA	RST	0.5886	0.63615	0.59207	0.58076	0.5755	0.59854
Gossipcop	PCA	LIWC	0.64683	0.71231	0.71101	0.66665	0.6643	0.73398
Gossipcop	PCA	RST_UPF	0.93199	0.94469	0.94467	0.90087	0.90645	0.95709
Gossipcop	PCA	LIWC_UPF	0.90764	0.93814	0.9344	0.90531	0.90343	0.95241
Gossipcop	PCA	UFS (ALL)	0.931	0.94689	0.94815	0.92293	0.92232	0.95943

Our extensive research consistently showed this dimensionality reduction technique's practical benefits. Machine learning models consistently displayed competitive accuracy after being trained on these lower-dimensional datasets, showcasing their superior generalization abilities.

Next, we trained and evaluate different machine learning models on our dataset. A thorough assessment of several machine learning models like bagging Classifier, Random Forest, Extra tree Classifier, Adaboost Classifier, Gradient Boosting, XBoost classifier using a variety of datasets and feature sets is shown in the table 2 that is provided. The datasets "Polifact" and "Gossipcop," which stand in for the data sources being analyzed, are included in the table 2 along with three distinct forms, "OD (original dataset)," "WRST (Wilcoxoxn rank sum test)," and "PCA (principal computer analysis)," which most likely correspond to various data processing methods. The sets of attributes used in the analysis are listed in the "Features" column. These sets include "UPF," "RST," "LIWC," "RST_UPF," "LIWC_UPF," and "ALL."

The numerical values in the table are performance scores, which are likely intended to evaluate the predictive or accuracy of the models. The outcomes show how the model performs differently for various datasets and feature combinations, offering important information about which algorithm and feature set works best for a particular dataset. This thorough assessment adds to the paper's scientific rigor and robustness by assisting in the selection of the best model and feature configuration for the study's goals. It is clear that for the two datasets under investigation, particular feature sets and algorithms produced the highest accuracy scores.

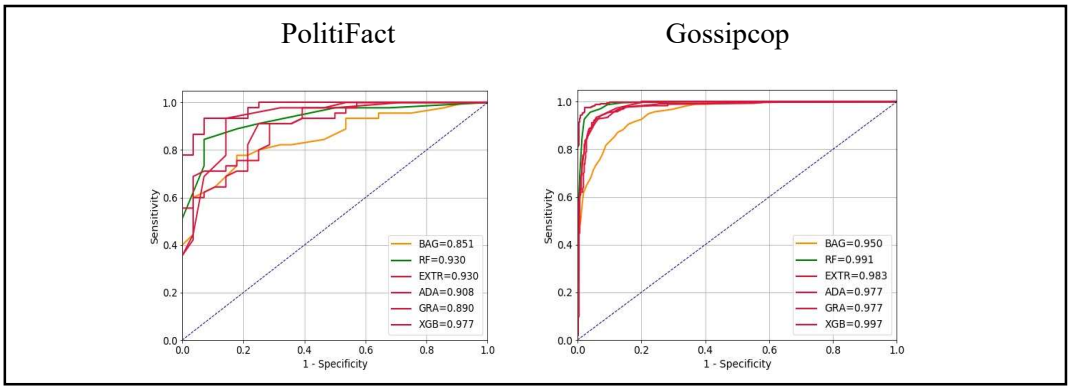

Figure 3: ROC Curve representation with Original Dataset-OD on UFS (ALL)

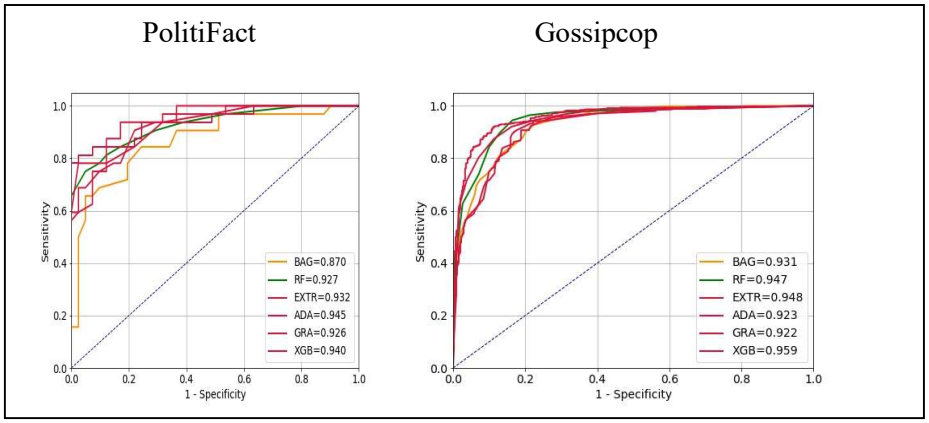

Figure 4: ROC Curve representation with Wilcoxon Rank Sum test-WRST on UFS (ALL)

Journal of Theoretical and Applied Information Technology
15th January 2024. Vol.102. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

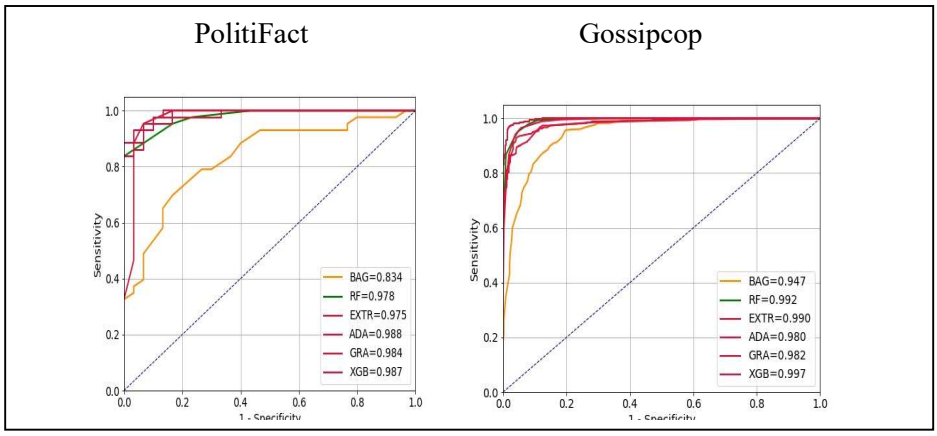

Figure 5: ROC Curve representation with Principal Component Analysis on UFS (ALL)

The "ALL" feature set produced the most accurate results for the "Polifact" dataset, with an accuracy of 0.978979. On the other hand, the "ALL" feature set performed better than the others with an astounding accuracy of 0.99791 for the "Gossipcop" with original dataset.

These results highlight the significance of feature engineering and model selection in the context of the research objectives, highlighting the need to choose the right feature set and algorithm combination to maximize predictive accuracy.

The Receiver Operating Characteristic (ROC) diagram is an essential tool in our research paper for evaluating the effectiveness of the XGBoost classification model. With specificity (True Negative Rate) on the horizontal axis and sensitivity (True Positive Rate) on the vertical, the ROC curve offers a graphical representation. Three different datasets were evaluated: Figure 3 shows the original dataset, Figure 4 shows a dataset that was subjected to the Wilcoxon Rank Sum Test, and Figure 5 shows a dataset that was subjected to Principal Component Analysis (PCA). The three datasets accuracy results highlight the XGBoost model's versatility and resilience, demonstrating its ability to classify positive and negative examples in a range of data scenarios.

Table III: Performance Metrics of Machine Learning Models on Fake News Detection Datasets

Dataset	Model	Accuracy	Precision	Recall	F1-measure
Politifact	Bagging	87%	86.55	86%	86%
	RF	92.7%	92.2%	92%	91%
	Extra Tree	93.22	93.11	93%	92.87
	AdaBoost	94.5%	94%	94.77	93%
	Gradient Boosting	92.6%	92%	92%	91%
	XGboost	**94.0%**	**93.87**	**92.88**	**93%**
Gossipofact	Bagging	93.1%	93%	93%	92.88%
	RF	94.7%	94.5%	94.3%	94%
	Extra Tree	94.8%	94.55%	94.00	93.97%
	AdaBoost	92.3%	92%	92%	92%
	Gradient Boosting	92.22%	92%	92.88%	91%
	XGboost	**95.99%**	**95.55%**	**95.11%**	**95%**

We compared their performance in terms of accuracy, precision, recall, F1 score, and other relevant metrics to determine which model works best for our specific problem and XGboost has achieved the highest accuracy our proposed Unified feature set(ALL). The outcomes of our experiment, which assessed the effectiveness of different machine learning models on the "Politifact" and "Gossipofact" datasets, are shown in Table 3. Evaluating these models' performance

for the classification task in terms of accuracy, precision, recall, and the F1-measure was the goal.

We found that the "Bagging" model obtained 87% accuracy, 86.55% precision, 86% recall, and 86% F1-measure in the "Politifact" dataset. Comparably, the "Random Forest (RF)" model showed 92.7% accuracy, 92.2% precision, 92% recall, and 91% F1-measure. The performance metrics of various models, including "Extra Tree," "AdaBoost," "Gradient Boosting," and "XGboost," were also presented on this dataset.The "XGboost" model on the "Gossipofact" dataset.

5. CONCLUSION

This study introduces new methods and clarifies important issues, marking a substantial advancement in the field of fake news detection. The importance of a holistic data representation has been demonstrated by the thorough feature integration of linguistic features (LIWC), user profile features (UPF), and Rhetorical Structure Theory (RST) features. It emphasizes how crucial it is to comprehend the subtleties of language, user behavior, and structural traits in order to recognize fake news with accuracy. Furthermore, one significant challenge in high-dimensional datasets is addressed by using Principal Component Analysis (PCA) for dimensionality reduction. It makes the models for detecting fake news more reliable and robust by protecting against overfitting and enhancing computational efficiency. The thorough model evaluation highlights the adaptability and versatility of fake news detection systems. It includes a range of machine learning models, "Politifact" and "Gossipofact" datasets, and diverse feature sets. This thorough investigation helps to understand the variables affecting model choice, offering vital information for developing and putting into practise effective fake news detection systems. The importance of selecting the right model and features is especially highlighted by the finding that the XGBoost model performs best on the unified feature set (ALL). This conclusion highlights how important it is to choose the best combination in order to maximize predictive accuracy—a crucial factor when it comes to detecting fake news.

The potential application of this research is quite promising. There is room for improvement in the form of investigating more sophisticated machine learning models, investigating different feature categories, and broadening the scope of the research to include a wider range of datasets. The XGBoost classification model's performance is visualized through the ROC curve analysis, which provides a comprehensive understanding of the model's abilities to classify positive and negative examples in a range of scenarios. This enhances the detection process's comprehensibility and transparency, promoting confidence in the system's decision-making. Future work would delve into refining models for broader applicability, ensuring interpretability, and addressing the evolving landscape of misinformation detection.

6. ACKNOWLEDGEMENT

This research was independently funded without any financial assistance from public, private, or nonprofit entities. We would like to express our heartfelt appreciation to the faculty members of the Department of Computer Science Engineering at Christ University.

REFERENCES:

[1] H. Allcott and M. Gentzkow, “Social Media and Fake News in the 2016 Election,” Journal of Economic Perspectives, vol. 31, no. 2, pp. 211–236, May 2017, doi: 10.1257/jep.31.2.211.

[2] Xinyi Zhou and Reza Zafarani.” 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities”. ACM Comput. Surv. 53, 5 (2020),1–40. https://doi.org/10.1145/3395046

[3] M. C. Arcuri, G. Gandolfi, and I. Russo, “Does fake news impact stock returns? Evidence from US and EU stock markets,” Journal of Economics and Business, p. 106130, Jul. 2023, doi: 10.1016/j.jeconbus.2023.106130.

[4] Kang, C. and A. Goldman. 2016.” In washington pizzeria attack, fake news brought real guns.”The New York Times.

[5] Karami, Mansooreh & H. Nazer, Tahora & Liu, Huan. 2021. Profiling Fake News Spreaders on Social Media through Psychological and Motivational Factors. 225-230. 10.1145/3465336.3475097.

[6] Shu, K., D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint

arXiv:1809.012868.
[7] Shu, K., A. Sliva, S. Wang, J. Tang, and H. Liu. 2017. Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter, 19(1):22–36
[8] Jin, Z., J. Cao, Y. Zhang, and J. Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In Thirtieth AAAI conference on artificial intelligence.
[9] L. Allein, M.-F. Moens, and D. Perrotta, "Preventing profiling for ethical fake news detection," Information Processing & Management, vol. 60, no. 2, p. 103206, Mar. 2023, doi: 10.1016/j.ipm.2022.103206.
[10] H. Tang, B. Liu, and J. Qian, "Content-based and knowledge graph-based paper recommendation: Exploring user preferences with the knowledge graphs for scientific paper recommendation," Concurrency and Computation: Practice and Experience, vol. 33, no. 13, Feb. 2021, doi: 10.1002/cpe.6227.
[11] Ryoya Furukawa, Daiki Ito, Yuta Takata, Hiroshi Kumagai, Masaki Kamizono, Yoshiaki Shiraishi, and Masakatu Morii. 2022. Fake News Detection via Biased User Profiles in Social Networking Sites. In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '21). Association for Computing Machinery, New York, NY, USA, 136–145. https://doi.org/10.1145/3486622.3493939
[12] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media," Big Data, vol. 8, no. 3, pp. 171–188, Jun. 2020, doi: 10.1089/big.2020.0062.
[13] C. I. Eke, A. A. Norman, L. Shuib, and H. F. Nweke, "A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions," IEEE Access, vol. 7, pp. 144907–144924, 2019, doi: 10.1109/access.2019.2944243.
[14] X. Dong, T. Li, R. Song, and Z. Ding, "Profiling users via their reviews: an extended systematic mapping study," Software and Systems Modeling, vol. 20, no. 1, pp. 49–69, Mar. 2020, doi: 10.1007/s10270-020-00790-w.
[15] K. G. Gatziolis, N. D. Tselikas, and I. D. Moscholios, "Adaptive User Profiling in E-Commerce and Administration of Public Services," Future Internet, vol. 14, no. 5, p. 144, May 2022, doi: 10.3390/fi14050144.
[16] A. Zhang, D. Hammer, A. Brookhouse, F. Spezzano, and L. Babinkostova, "Predicting the Influence of Fake and Real News Spreaders in Twitter," SSRN Electronic Journal, 2022, Published, doi: 10.2139/ssrn.4201848.
[17] A. Shrestha and F. Spezzano, "Characterizing and predicting fake news spreaders in social networks," International Journal of Data Science and Analytics, vol. 13, no. 4, pp. 385–398, Nov. 2021, doi: 10.1007/s41060-021-00291-z.
[18] S. V. Balshetwar, A. RS, and D. J. R, "Fake news detection in social media based on sentiment analysis using classifier techniques," Multimedia Tools and Applications, vol. 82, no. 23, pp. 35781–35811, Mar. 2023, doi: 10.1007/s11042-023-14883-3.
[19] S. Kh. Hamed, M. J. Ab Aziz, and M. R. Yaakub, "Fake News Detection Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments," Sensors, vol. 23, no. 4, p. 1748, Feb. 2023, doi: 10.3390/s23041748.
[20] G. Zhang, A. Giachanou, and P. Rosso, "SceneFND: Multimodal fake news detection by modelling scene context information," Journal of Information Science, p. 016555152210876, Apr. 2022, doi: 10.1177/01655515221087683.
[21] F. Celli and M. Poesio, "Pr2: A language independent unsupervised toolfor personality recognition from text," arXiv preprint arXiv:1402.2796, 2014.
[22] Y. Ji and J. Eisenstein, "Representation learning for text-level discourse parsing," in ACL'2014, vol. 1, 2014, pp. 13–24.
[23] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.
[24] K. Shu, X. Zhou, S. Wang, R. Zafarani and H. Liu, "The Role of User Profiles for Fake News Detection," in 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Vancouver, BC, Canada, 2019 pp. 436-439. doi: 10.1145/3341161.3342927