**Ensemble Learning Approaches for Identifying Fake news in Urdu Text**

**ABSTRACT**

The growing usage of social media makes the spread of fake news and misleading information easier. As a result, noting and detecting phony news has become an essential domain of study. While several techniques have been proposed for identifying real news from fake, they frequently prove ineffective when applied to various datasets, especially those written in Urdu. Furthermore, some studies use machine translation datasets from Google Translate or similar platforms instead of human review, making it more difficult to identify patterns in English news translated into Urdu accurately. This restricts the use of these methods in practical settings. This paper proposes the classification of false news as Urdu news as a solution to these problems. This dataset comprises three thousand two hundred eighty-four news items total from nine distinct domains. Bow's bag of words and the phrase frequency-inverse document frequency (TF-IDF) were used in the process. The main contribution of this study is its practical application in stacking, random search, and grid search, which integrate verb extraction from the preprocessed text with feature vectors of the preprocessed text. In addition to logistic regression and support vector machines, ensemble techniques such as random forests (RF) and extra trees (ET) are also used for bagging. Boosting techniques such as gradient boosting and XGBoost were also used. However, stacking worked well since logistic regression was a meta-learner, and extra trees, gradient boosting, and random forest served as base learners. According to experimental data, stacking achieves accuracy, specificity, sensitivity, MCC, ROC, and F1 scores of 95.28%, 95.89%, 94.72%, 90.57%, 98.97%, and 95.00%, respectively.

**INTRODUCTION**

Due to its extensive usage, social media has had a significant impact on individuals, particularly in terms of the rapid dissemination of news. As per the Pew Research Report 2021, 86% of people in the United States rely on the internet as their primary news source, surpassing traditional news channels [1]. With social media's widespread availability and global reach, spreading fake, manipulated, or fabricated content has become more convenient than ever. As a result, people and institutions work to mislead and control the public by providing them with the information they want. After growing in popularity online, Collins Dictionary declared the term "fake news" the word of the year in 2017 [2]. Fake news can damage people's perceptions of governments, businesses, organizations, and individuals, as well as their reputations and goodwill. It may also result in unrest and financial hardship. Because manual identification would be expensive and wasteful, automated methods and approaches are thus required to identify fake news [3]. The purpose of fake news is to intentionally confuse people and damage the public's trust [4]. To address this issue, creating an automated tool that can verify the credibility or falsity of news is imperative. Natural language processing (NLP) is a methodology used to manage textual information, allowing computers to comprehend and engage with human language [5]. Many applications, including text classification, query answers, and language translation, can be made possible using NLP [6]. The most common part of NLP challenges is text classification, which focuses on figuring out a word's, phrase's, or document's semantic meaning [7]. Detecting fake news is a challenging task with many challenges. The tedious process of creating benchmark datasets and manual annotation are two examples of challenges faced. This problem is even worse when using a resource-poor language, such as Urdu, which does not have many internet resources. There are many techniques for finding fake news in Urdu, but they often have drawbacks. These techniques typically do not use data from multiple domains. Incorporating data from many disciplines allows for a more thorough evaluation of predictive performance. We recommend using data from a broader domain to evaluate model performance.

[8] A study was conducted to find false information in five different areas. We used his Google Translate for English to Urdu translation during dataset collection for some studies. However, the actual application of such models is limited due to the lack of manual verification. Furthermore, the data set used had a small number of samples, which could have prevented proper training and testing of the model. An earlier study that examined fake news from five domains included just 900 items in its dataset.

This study focuses on these issues in identifying fake Urdu news. To identify false news from multi-domain data, a fake news detector is developed in this paper.

An ensemble classifier is suggested to identify fake news in Urdu. As base learners, use XGBoost, gradient boosting, and extra tree classifiers. As a meta-learner, use random forest.

The verb feature vector obtained from the preprocessed text was combined with the preprocessed text to complete feature stacking. To the authors' knowledge, feature stacking has not been applied to Urdu. A detailed collection of 3,284 news items from nine different domains was collected. The dataset is publicly available and manually labeled.

This study uses Vocabulary (Bow) and Frequency-Inverse Document Frequency (TF-IDF) techniques. This study includes experiments using Support Vector Machines (SVM), Logistic Regression (LR), Random Forests (RF), Extra Trees (ET), Gradient Boosting, and XGBoost.

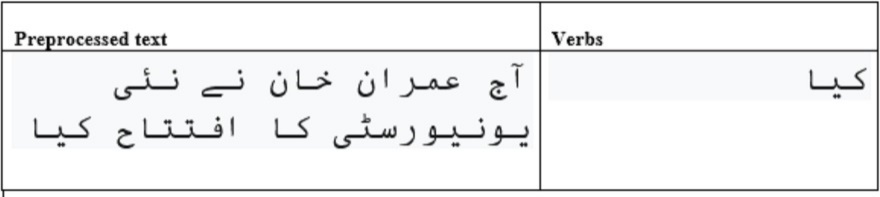


Figure 1 The process of extracting verbs from preprocessed text.

Feature stacking begins by generating the feature vector using the preprocessed text. Second, verbs are taken from the cleaned text, and then the verbs' feature vector is calculated. Figure 1 illustrates an example of feature extraction for the verb. Ultimately, these two feature vectors are combined into a single feature vector.

The remaining portion of this study is structured in the following manner. The section titled "Related Work" discusses prior works on fake news identification in the Urdu language. The section titled 'Materials and Methods' provides a detailed description of the suggested approach for detecting fake news, including the architectural design and an explanation of the procedure. Subsequently, there is an examination of the empirical findings. Ultimately, the conclusion is provided.

**RELATED WORK**

Various machine-learning models have been developed to detect counterfeit data in the Urdu language. Maaz Amjad completed two investigations to detect fake data in the Urdu language.

The benchmark dataset of Amjad et al. (2020) has 900 samples, with 500 samples representing real Urdu news and 400 samples representing fake Urdu news. However, due to the wide range of fields and the limited number of samples, this dataset is restricted. The model's capacity to detect fraudulent information in Urdu should be restricted due to the potential for real-time news arriving from domains with low data samples. One additional constraint of this study is that the researchers make fake news pieces, which may not be the most efficient method.

Amjad, Sidorov, and Zhila (2020) did a supplementary investigation to identify deceptive material in the Urdu language. This work produced numerous datasets by employing machine translation, enhancing data reduction techniques, and enlarging the initial dataset, as well as integrating a benchmark dataset. However, the main aim of the study was to assess different approaches rather than discovering new findings.

Akhter et al. (2021) utilized an ensemble methodology to detect fraudulent news in the Urdu language. This research has been improved by incorporating a unique dataset and utilizing the benchmark dataset created by Amjad, Sidorov, and Zhila (2020). Regrettably, this undertaking is unable of producing a highly precise model for detecting fraudulent news in the Urdu language. In addition, this study utilized Google Translate to translate an English-to-Urdu dataset, eliminating the requirement for human verification. This study has a narrow focus, as it solely incorporates data from a mere five domains.

Lina, Fua, and Jianga proposed a complex deep learning model called CharCNN-RoBERT in their 2020 study. The algorithm was designed specifically to detect false news items written in Urdu. This research employs a grand total of 900 samples extracted from the dataset supplied by Amjad et al. (2020). The study utilized pre-training, charCNN, Roberta, as well as word and character n-grams to detect fraudulent news in the Urdu language. The findings illustrate that by integrating pre-training, label smoothing, charCNN, and Roberta, it is feasible to achieve a precision of 0.90 in detecting misinformation in the Urdu language. Balouchzahi and Shashirekha (2020) utilized machine learning methods to develop models capable of identifying deceptive content in the Urdu language. The models were trained using word and character n-grams. Additionally, word embedding vectors are utilized in the training stage of deep learning models to achieve a comparable result. The ensemble technique based on machine learning has above-average performance, with an F1 score of 0.78 and an accuracy of 0.79.

The research presented above are constrained by several constraints. Initially, the dataset was translated from English to Urdu without passing manual review. Moreover, the dataset has a restricted number of samples, which hinders the effective evaluation of the models. Prior research was limited to a smaller number of areas, resulting in insufficient representation across many aspects.

This study on fake news detection in Urdu presents a comprehensive analysis of a dataset comprising 4,097 news articles collected from nine diverse domains. The dataset underwent meticulous manual verification to ensure the accuracy of labeling between real and fake news. To detect phony news effectively, the (Muhammad Shoaib Farooq 2023) study employed an ensemble classifier combining Random Forest, Extra Trees, and Logistic Regression models. Feature engineering techniques such as term frequency-inverse document frequency (TF-IDF) and bag-of-words (Bow) with word and character-level n-grams were utilized to enhance the performance of the models. The experimental results showcased the robustness of the proposed approach, achieving an impressive accuracy rate of 93.39%. This research contributes significantly to the field of fake news detection in Urdu, offering valuable insights for improving the accuracy and reliability of news verification processes in the digital era. In contrast, this (Muhammad Shoaib Farooq 2023) study includes data acquisition from nine distinct domains, enabling a more comprehensive analysis across various dimensions.

An in-depth overview of the studies that are being discussed.

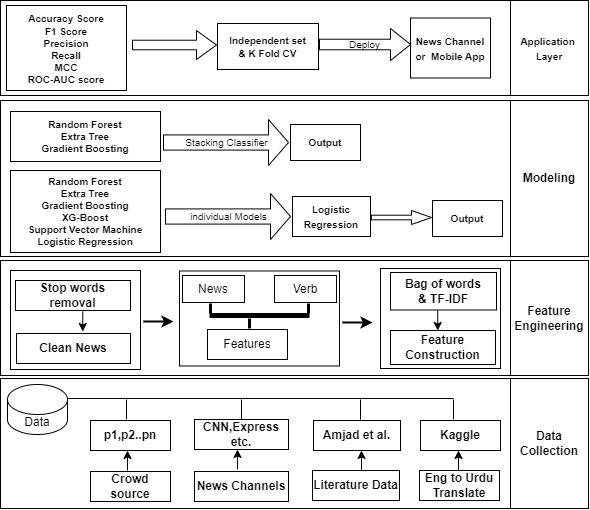
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Research | **Features** | **models** | **Dataset** | **Problems** | **Domains** | **Machine Translate Manually verified** |
| Amjad et al. (2020) | Word, character, and functional n-grams | AdaBoost, LR, SVM, RF, MNB, BNB, DT | 900 | Small dataset 900 news. Fake news is created based on real news. | 5 | – |
| Amjad, Sidorov, and Zhila (2020) | Word, character, and functional n-grams | SVM, AdaBoost | 400 | Machines translate datasets. No manual verification. | – | No |
| Akhter et al. (2021) | Character trigrams, | NB, DT,SVM | 2000 | Machines translate datasets. No manual verification-optimized results | – | No |
| Lina, Fua, and Jianga (2020) | Bow and Information Gain | RoBERTa, charCNN | 900 | No dataset is contributed. | – | – |
| (Muhammad Shoaib Farooq 2023) | Bow and tf-idf with n-grams | LR, SVM, RF, ET and stacking | 4097 | There are differences between datasets. Real news is lower to fake news. Compared to default settings, default models don't use grid search CV and random search CV to choose the optimal parameters and achieve high accuracy. This is a major problem. | 9 | Yes |

Table 1 provides an overview of the constraints encountered in the current research regarding the identification of counterfeit news in Urdu and the comparison of various studies. The objective of this study is to address these challenges associated with the detection of fake news in Urdu. This research introduces a novel dataset and a dedicated classifier specifically designed to identify false news in Urdu within a diverse dataset spanning multiple domains. Unlike previous studies that utilized 900 news instances from five domains, the dataset has been significantly expanded to include 4097 news from nine domains (Muhammad Shoaib Farooq 2023). Furthermore, the news obtained from Google translation has undergone manual verification, as it is not entirely reliable for accurate translation.

The dataset used in the study on detecting English fake news (Ahmed, Traore & Saad, 2017) was acquired from Kaggle. To address the issue, each news article underwent manual verification. Feature engineering techniques such as TF-IDF and (Bow)were employed in the process.

**Resources & Methods**

This section provides a detailed analysis of the architecture of the chosen technique. This section also discusses the ensemble approaches of bagging, boosting, and stacking. The chosen methodology for detecting false information is illustrated in Figure 2.



**Data Collection**

The initial layer involves gathering data from many sources such as news websites, previous studies, and datasets that have been translated by machines. The authentic news dataset was obtained from multiple sources, including a literature dataset from Amjad, Sidorov & Zhila (2020), as well as various news channels' websites such as BBC Urdu News, Dawn News, and City 42 News. The data was collected between May 2020 and March 2021. There is currently no adequate repository for false information in the Urdu language, which has limited resources. Various sources, including Vishvas’s news, were utilized to gather false news data in Urdu. The English fake news dataset was obtained and then manually verified to convert it into Urdu. Additionally, we acquired a dataset of Urdu fake news via a published article (Amjad, Sidorov & Zhila (2020) as well as through crowdsourcing.

**Data annotation strategy**

Three distinct annotators have been employed to categorize the acquired news. These individuals have been selected based on their specialized areas of knowledge to guarantee effective data gathering.

* Proficient in using social media platforms.
* Urdu native speakers.
* Possessing a master's degree or higher in the relevant field.
* Having expertise in data annotation.
* Excluded from all jobs on news or social media platforms.

**Real news gathering sources**

* Dawn News: [www.dawnnews.tv](http://www.dawnnews.tv/)
* Express Newspaper: [www.express.com.pk/](http://www.express.com.pk/)
* City 42 News: [www.city42.tv](http://www.city42.tv/)
* BBC Urdu News: [www.bbc.com/urdu](http://www.bbc.com/urdu)

**Real news acquisition**

The dataset has been manually annotated to identify real news, and other aspects have been considered to define the criterion for accepting real news.

According to a credible online source

A real news channel.

A real newspaper.

**Fake news acquisition**

Urdu is a language that lacks sufficient resources; there is no existing collection of Urdu false news. The Urdu fake news dataset has been compiled from three primary sources: (i) websites, (ii) crowdsourcing, and (iii) existing fake news datasets for the English language mentioned in a research study (Ahmed, Traore & Saad, 2017). The fake news data released in English has been translated into Urdu through human verification. If any news appears to be illogical, it is likely that the story has been either withdrawn or rectified by referring to the official source. Finally, the fraudulent news dataset was obtained from a research paper published by Amjad et al. (2020) specifically for the purpose of detecting fake news in Urdu. Table 3 provides more sources of fabricated news data.

**Sources of fake news collection**

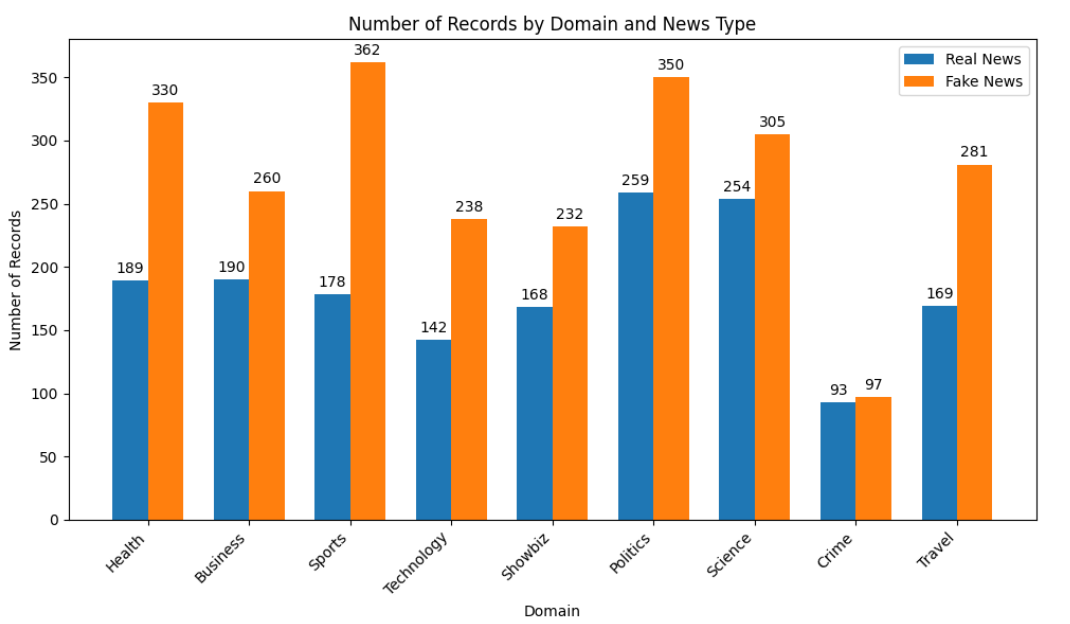
|  |  |
| --- | --- |
| **Sources of news** | **URL** |
| Urdu fake news dataset | <https://github.com/MaazAmjad/Datasets-for-Urdu-news> |
| Crowdsources | **---------------** |
| Vishvas’s news | [www.vishvasnews.com/urdu](http://www.vishvasnews.com/urdu) |
| Fake news dataset | <https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset> |

The crowd-sourced experts are instructed to create arbitrary fake news, resulting in an impartial collection of facts. Figure 3 illustrates an instance of a chosen strategy. In the earlier study conducted by Amjad et al. (2020), professionals sourced from the crowd were instructed to create fabricated news by making slight modifications to genuine news articles. This resulted in a dataset that was skewed or influenced by bias. This study, however, does not produce fabricated news and only examines those that are present in established datasets or acquired from other sources described in Table 3.

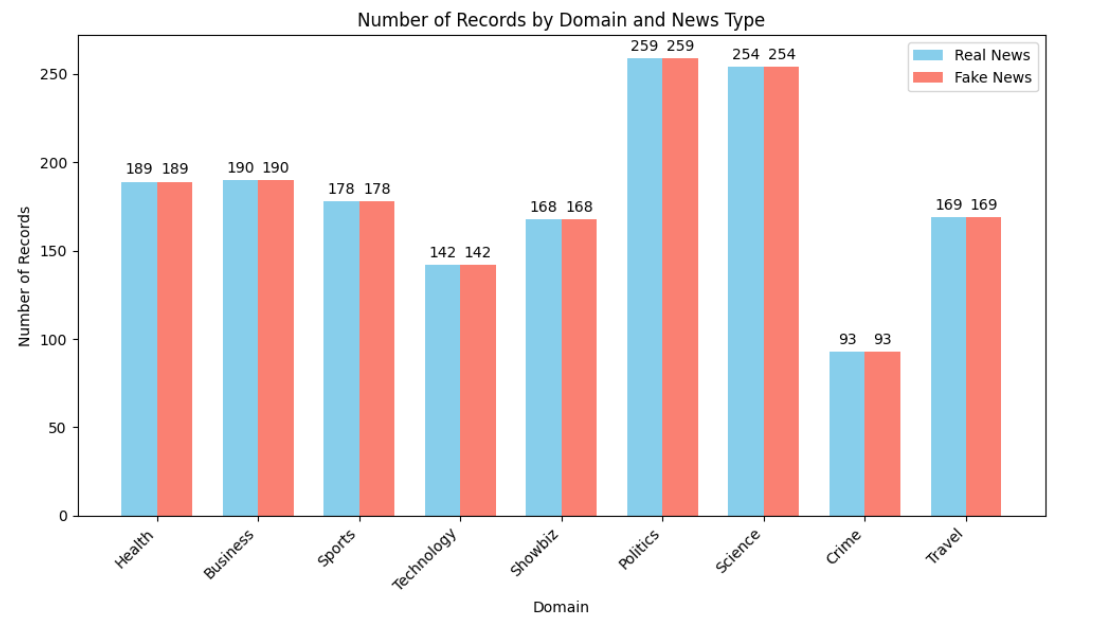
**Figure 3 Real *vs* fake news**

To assess the effectiveness of dataset annotation by three annotators, we calculated the inter-annotator agreement (IAA) using Cohen's Kappa coefficient. This statistical measure helps determine the reliability of two annotators. The assessment results in a 92% cumulative score.   
Finally, a dataset in the English language from Kaggle was chosen specifically for machine-translated news. The Instant Scrapper tool was utilized to gather authentic news data from many websites. The spurious news data was gathered from the English language, thereafter, transcribed into the Urdu language, and the translated news was meticulously verified (Ahmed, Traore & Saad, 2017). If any news fails to accurately convey its intended implications, it is either eliminated or manually corrected to ensure the proper sequence or language. Figure 4 depicts a discarded news sample.

Over the past three years, we gathered both counterfeit and authentic news via a range of methods and methodologies. The corpus has a total of 4,097 records, consisting of 2,455 fake news and 1,642 real news. While creating a corpus, researchers evaluated many sorts of material. However, the annotation method may differ differently. Prior to utilizing the dataset, it underwent a thorough verification process. Prior research has concentrated on certain domains, such as politics, and constructed models for individual datasets. This approach is susceptible to dataset biases and is likely to have poor performance when applied to news from a different domain. To address these problems, our study gathered news articles from nine distinct categories. Figure 4 displays the domains from which news articles have been gathered, together with the corresponding quantities of genuine and fabricated news within each domain.



In our investigation, we utilized a dataset that was developed by Usman. However, a significant issue in the dataset is the lack of balance, particularly in most fields where there is a higher prevalence of fake news compared to actual news.

**Feature Engineering Layer**

Due to the presence of noise and redundant information, the raw data was processed at the second layer to ensure its readiness. It diminishes the precision of the model and increases the processing duration. Prior to being entered into the system, the data must adhere to a defined and consistent learning model.

**Deletion of Stop Words**

Stop words facilitate the concentration of machine learning models on more meaningful words by limiting the amount of text data.

**Benefits:**

* To mitigate noise in the data, we eliminate stop words, hence facilitating the models' ability to discern crucial information.
* Removing these commonly utilized phrases from the model aids in enhancing memory throughout training.
* Machine learning models can achieve higher performance when they favor relevant keywords over confusing stop words.

**Features**

Features play a vital role in machine learning for text classification. For this investigation, we utilized unprocessed, purified text characteristics obtained from analyzed news articles, as well as verbs extracted as features. We extracted verbs as features from preprocessed news after initially computing features on the preprocessed news. Ultimately, these two attributes were utilized to generate feature vectors, which were subsequently merged to form a unified feature vector. The model was provided with this feature vector to determine the authenticity of news articles, distinguishing between those that are fake and those that are real.

**Feature construction**

Two primary approaches were utilized to calculate features, namely TF-IDF and Bag of Words (Bow). The TF-IDF statistical technique is utilized to ascertain the significance of a word within a text or corpus, as elucidated by Liu and colleagues (2018). The phrase frequency in news refers to the frequency of occurrence of a specific term. The importance of a term is evidenced by how often it appears in news articles. A matrix is created by analyzing the word frequency of each term in the news. The matrix is structured such that its rows correspond to the number of news items, while its columns correspond to the number of unique phrases. Document frequency measures the extent to which a particular characteristic is present in news items, therefore indicating its accessibility. The inverse document frequency (IDF) diminishes the significance of a characteristic if it is evenly distributed throughout all news stories. This is the procedure by which characteristics are assessed and given a comparative significance. The primary utilization of BOW entails the transformation of news articles into numerical representations. The suggested method employs a frequency-based count. The fixed-length matrix determines the frequency of a term by counting how many times it appears in each news report.

**Modeling**

At this level, a fake news classifier is created utilizing various machine learning models. The algorithms, such as bagging, boosting, and stacking, are collaborative. The text above provides a summary of the algorithms employed in each strategy. The methodologies section outlines the structure of various machine learning algorithms employed to enhance the accuracy of the flawed news classification system. This study included various machine learning approaches, including XGBoost, Extra Tree, Gradient Boosting, Random Forest, Support Vector Machine, and Logistic Regression. Each approach was examined individually to identify counterfeit Urdu information. Subsequently, more algorithms were used to enhance the precision of the models. Below, you can find a concise summary of the reasoning for each model.

**Logistic Regression**

Logistic regression is a commonly used statistical model for predicting the likelihood or probability of an event happening based on one or more input variables or characteristics in classification tasks. Logistic regression produces probabilities ranging from 0 to 1, unlike linear regression which provides continuous findings. This characteristic of logistic regression enhances comprehension and strengthens decision boundaries. The sigmoid function used to a linear combination of weights and input features in logistic regression can transform any real-valued number into a range of [0, 1]. This results in a seamless curve that resembles the shape of the letter S. Maximum likelihood estimation is used to determine coefficients by optimizing the alignment between forecasts and actual observations. These coefficients provide professionals with valuable information about the relative significance of different features, enabling them to determine the main factors contributing to observed events.

**Support vector machine**

After acquiring the input, the SVM classifier constructs the most effective separation hyperplane. The decision border is the delineation that separates news items into discrete groups. It is essential to determine the decision boundary of the linear kernel to choose the suitable model (Alexandre 2015). The boundary line was determined using the subsequent equation. The variable x1 represents the news feature, ω1 indicates the coefficient weight, and b represents the slope.

y = b + ω1 ∗ x1 + ω2 ∗ x2+... **(1)**

After establishing the hyperplane, we can employ it to make predictions. The underlying principle of the function h is expressed as follows.

**Add one more equation (2)**

When a new instance is created, its score is classified as positive or negative depending on whether it is greater than or equal to zero. Support Vector Machines (SVM) can accommodate many hyperplanes. Enhancing the separation between the hyperplanes of the two classes can lead to an improvement in the performance of Support Vector Machines (SVM).

**Random forest**

RF, a member of the bagging family, utilizes bootstrapping to distribute samples based on Attique et al.'s (2020) methodology. The method uses replacement sampling to build sub-datasets of authentic and synthetic news articles. The distribution of spurious and genuine news is evenly balanced across all subgroups. The model is provided with the news as a feature vector, along with a label, for training. Decision trees are formed by randomly selecting the best-split nodes. Every weak learner is allocated a test instance, and the class prediction is established by a majority vote.

**Extra tree**

Another bagging method that trains models without replenishment and includes both the news and the label is referred to as ET. When the reports in each subgroup have equal sizes, ET generates numerous sub-datasets. Upon input into the model, each weak learner is provided with a query instance, and the class prediction is determined by the majority vote.

**Gradient boosting**

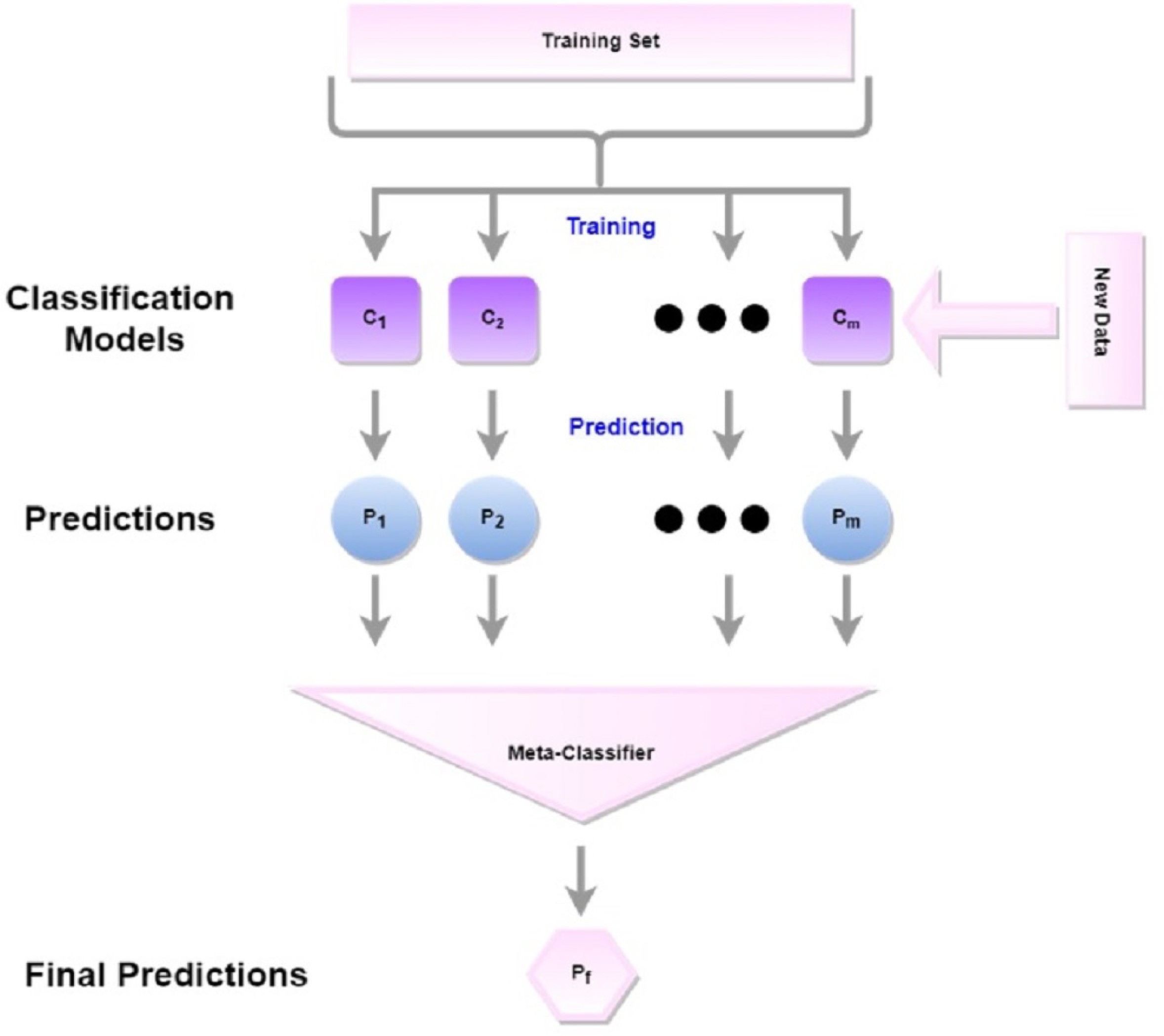
Gradient boosting is a widely used machine learning approach for solving regression and classification issues. The process entails amalgamating the most optimal following model with previous models to decrease the occurrence of prediction errors. The primary concept is to establish target outcomes for the succeeding model to minimize the overall error. The ultimate forecasts are produced using a Gradient Boosting Machine, which amalgamates the predictions from numerous decision trees. It is crucial to emphasize that decision trees serve as weak learners in a gradient-boosting machine.

**XGBoost**

XGBoost, also known as Extreme Gradient Boosting, is a machine learning technique that belongs to the ensemble learning category. It is frequently employed for supervised learning applications, such as regression and classification. XGBoost develops a predictive model by amalgamating the forecasts of many models, usually decision trees. To maximize its efficacy, the ensemble incorporates weak learners, each of which is dedicated to correcting the mistakes produced by the preceding learners. XGBoost employs gradient descent to minimize a pre-defined loss function during the training procedure.

**Stacking**

According to Figure 5, stacking is an ensemble strategy where classifiers provide predictions for each news item using multiple models and combine these predictions to create a new dataset. The three highest-performing classifiers, namely additional tree, gradient boosting, and XGBoost, have been chosen as basic learners. A 2D vector is created by the forecasts generated by the additional tree, gradient boosting, and XGBoost algorithms for each news item. The final learner utilizes this 2D vector and divides it into separate training and testing sets.



When the weights are provided to the model as a feature vector, the input data is multiplied by them. After computing the aggregate of all news sources, the resulting sum is sent via a sigmoid function. The sigmoid function restricts the probability range to values between 0 and 1 by employing a threshold of 0.5 (Zhang, Yang, & Zhang, 2018; Tolles & Meurer, 2016). The sigmoid function is denoted by Y in the given formula, whereas Z represents the sum of all the integers.

**z = (w1x1 + w2x2+,...,wnxn + b** (3)

Add one more equation (4)

**Machine learning model parameters**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Model** | **Hyperparameters** |
| 1 | Random Forest | n\_estimators: 300, max\_depth: None, random\_state: 42 |
| 2 | Extra Trees | n\_estimators: 550, max\_depth: None, random\_state: 42 |
| 3 | Gradient Boosting | n\_estimators: 250, learning rate: 0.2, max\_depth: 7, random\_state: 42 |
| 4 | XGBoost | n\_estimators: 300, max\_depth: 3, random\_state: 42 |
| 5 | Logistic Regression | C: 100, max\_iter: 100, penalty: ‘l2’, solver: ‘liblinear’, random\_state: 42 |
| 6 | Support Vector Machine (SVM) | C: 1.0, kernel: ‘rbf’, gamma: ‘scale’, probability: True, random\_state: 42 |
| 7 | Stacking | Base estimators: Extra Trees, Gradient Boosting, and XGBoost, Final estimator: RandomForestClassifier, random\_state: 42 |

The hyperparameters for seven distinct machine learning models are listed in table 2 above. The performance of the models is heavily influenced by these hyperparameters. The parameters for tree-based models (Random Forest, Extra Trees, Gradient Boosting, XGBoost) include 'n\_estimators', 'max\_depth', and 'random\_state'. For Logistic Regression, the parameters are 'C', 'max\_iter', 'penalty', and 'solver'. For SVM, the parameters are 'C', 'kernel', 'gamma', and 'probability'. Lastly, for Stacking, the parameters are 'Base estimators' and 'Final estimator'. The 'random\_state' parameter is universally shared among all models to guarantee the replicability of outcomes. It is crucial to acknowledge that these hyperparameters are specific to the model and their optimal values can differ depending on the data and the specific problem being addressed.

**Application Layer**

The application layer establishes a connection between mobile apps and web platforms by incorporating models. It aligns with the prevailing mobile-first trend in technology by enhancing accessibility and optimizing user interaction for social media and news consumption.

**Evaluation parameters**

To assess the performance of the model, several evaluation metrics were utilized. The study utilizes metrics such as specificity, sensitivity, accuracy, and Mathew's correlation coefficient.

The accuracy score of a model is a crucial metric to measure its performance. The accuracy formula is shown below.

Accuracy = True Positive + True Negative

True Positive + True Negative + False Positive + False Negative (5)

Specificity was used to give a numerical evaluation of the model's predictive accuracy in negative situations. The formula for specificity is provided below.

Specificity = True Negative (6)

True Negative + False Positive

The model's capacity to recognize positive class samples is measured by sensitivity, also known as recall. Here is how to calculate sensitivity:

**Sensitivity** = True Positive (7)

False Positive + False Negative

The F1 score is used to measure the accuracy of a classifier by considering both precision and recall. The F1 score evaluates the performance of a classifier by considering the number of true positives, false positives, and negatives, resulting in a balanced rating. A high F1 score signifies excellent precision and recall, whereas a low score indicates subpar performance.

(8)

**The Matthews correlation coefficient (MCC) is a robust and consistent statistical tool. It assesses prediction performance across all four quadrants of the confusion matrix. A high MCC score indicates strong performance. The formula for MCC is as follows:**

Add F1 formula (9)

The ROC AUC value measures the classifier's capacity to differentiate between positive and negative classifications. The ROC curve provides information by plotting the true positive rate versus the false positive rate. A classifier with a high ROC AUC value is regarded as very accurate in distinguishing between positive and negative cases.

**Experiments**

Here is a brief description of the dataset, emphasizing the number of words and the range of vocabulary:

* The **pure corpus** contains **316,355** words.
* There are **23,222** unique words in the dataset.

**Results**

The model's performance and robustness are evaluated through two sets of experiments. The findings of each of these studies will be examined and discussed in this section.

**Independent set testing**

An established method for assessing the classifier's performance using concealed data is independent set testing. In this examination, the data is often categorized into two distinct groups. The initial segment pertains to the training set, comprising input and output pairs that are provided to the model to facilitate effective learning. The second half comprises a test set in which just input features are provided, while labels are concealed. The model must accurately classify instances as either false or real based on the provided attributes. The results of the independent set testing are displayed in Table 8.

**Results with TF-IDF in independent settings**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sr. No | Model | Accuracy | Specificity | Sensitivity | MCC | F1 Score |
| 1 | Random Forest | 94.97% | 95.88% | 94.13% | 89.96% | 95.11% |
| 2 | Extra Tree | 94.52% | 95.56% | 93.54% | 89.06% | 94.65% |
| 3 | Gradient Boosting | 92.99% | 95.56% | 90.61% | 86.13% | 93.07% |
| 4 | XGBoost | 92.99% | 93.35% | 90.61% | 83.91% | 92.10% |
| 5 | Logistic Regression | 91.93% | 93.35% | 90.61% | 83.91% | 92.10% |
| 6 | SVM | 90.71% | 93.03% | 88.56% | 81.54% | 90.82% |
| 7 | Stacking Classifier | 95.28% | 95.88% | 94.72% | 90.56% | 95.42% |

Table 8 displays the empirical outcomes obtained by employing TF-IDF characteristics. The results demonstrate that all models exhibit satisfactory performance, except for SVM, which achieves an accuracy of 90.71%. The stacked-based strategy has demonstrated the highest degree of reliability and Matthews Correlation Coefficient (MCC) score, reaching 95.28% and 90.56% respectively, in detecting fake and true news in Urdu. The models RF, ET, GB, XGBoost, and LR achieved MCC scores of 89.96%, 89.06%, 86.13%, 83.91%, and 83.91%, respectively.

Figure 6 displays the performance of the models in relation to the ROC-AUC curve. The SVM and LR models had the lowest ROC-AUC curve, whereas the proposed stacking model, as well as other models like bagging and boosting, achieved the highest ROC-AUC curve utilizing TF-IDF features.

**A graph of a function

Description automatically generated with medium confidence**

**Results with BOW in independent settings**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sr. No | Model | Accuracy | Specificity | Sensitivity | MCC | F1 Score |
| 1 | Random Forest | 93.60% | 93.98% | 93.25% | 87.20% | 93.60% |
| 2 | Extra Tree | 93.60% | 94.93% | 92.37% | 87.25% | 93.60% |
| 3 | Gradient Boosting | 90.41% | 94.62% | 86.51% | 81.17% | 90.40% |
| 4 | XGBoost | 90.32% | 91.77% | 89.73% | 81.45% | 90.51% |
| 5 | Logistic Regression | 90.71% | 91.77% | 89.73% | 81.45% | 91.00% |
| 6 | SVM | 89.95% | 92.72% | 87.39% | 80.07% | 90.00% |
| 7 | Stacking Classifier | 94.06% | 94.62% | 93.54% | 88.12% | 94.00% |

The findings obtained from machine learning models utilizing the Bag-of-Words (Bow) features are presented in Table 9. The results suggest that the models' performance is significantly diminished while utilizing the Bag-of-Words (Bow) features. The Bag of Words (Bow) method quantifies the frequency of terms in the corpus but does not take into account the significance of uncommon terms. While Bag-of-Words (Bow) models frequently yielded superior outcomes compared to more intricate models, the TF-IDF approach, which captures the significance of key terms, demonstrated greater performance in this investigation for detecting fake news in Urdu. Among the models that are being used, the stacked model generally demonstrates superior performance.

A red and blue squares with numbers and a blue bar

Description automatically generated

The confusion matrix in Figure 8A displays the performance of the best model for independent set testing. This model utilizes TF-IDF feature vectors, and the stacking model has demonstrated superior performance compared to other classifiers.

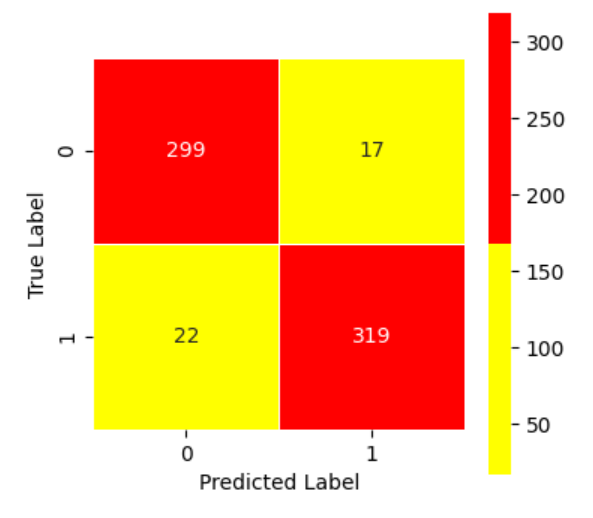


Figure 8B displays the optimal outcomes obtained for independent set testing with the Bag-of-Words (Bow) feature vector. The confusion matrix is specifically used to evaluate the performance of the layered technique, as it provides a clear representation of the most optimal outcomes.

For corroborating the performance of the proposed approach, the results of the proposed approach are compared with state-of-the-art existing models for Urdu fake news detection.

To ensure a just comparison, we applied the models from previous studies (Amjad, Sidorov & Zhila, 2020; Amjad et al., 2020; Akhter et al., 2021; Muhammad Shoaib) using the dataset obtained in our current study, and then we compared their performance. The performance comparison findings are presented in Table 12. The results indicate that the proposed approach demonstrates significantly superior outcomes when compared to previous research. We evaluated the model using the Matthews correlation coefficient (MCC), a robust measure that considers both groups. The suggested model demonstrates a superior accuracy of 95.28% in comparison to the previous leading study, which achieved an accuracy of 93.82%.

**Performance comparison with existing approaches**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Ref.** | **Accuracy** | **Specificity** | **Sensitivity** | **MCC** | **ROC** | **F1 score** |
| 1 | 84.23 | 93.20 | 82.30 | 75.34 | 0.940 | 84.30 |
| 2 | 83.48 | 91.34 | 83.21 | 74.38 | 0.956 | 87.37 |
| 3 | 81.85 | 89.33 | 84.37 | 70.37 | 0.927 | 82.72 |
| 4 | 93.82 | 88.96 | 96.33 | 86.20 | 0.983 | 93.17 |
| Current study | 95.28 | 95.88 | 94.72 | 90.56 | 0.989 | 95.42 |

**Arguments**The latest research in detecting Urdu fake news utilizes a reduced number of domains and datasets. These models are unreliable since the translations of Google English news are not manually reviewed. The dataset from the previous study conducted by Muhammad Shoaib Farooq in 2023 has experienced substantial growth. The collection has 4,097 news items encompassing various topics such as sports, business, technology, health, celebrities, politics, science, crime, and travel. This strategy specifically targets the challenges associated with imbalanced datasets in Table 4. We implemented various techniques to reduce the dataset's size, leading to an equitable representation of both domains. The findings are displayed in Table 6. In 2023, Muhammad Shoaib Farooq did a study which revealed that Urdu language algorithms employ uneven default settings, resulting in a detrimental effect on model performance and the emergence of biases. No hyperparameter optimization using grid search or random search was performed for the default configuration on an imbalanced dataset. All machine learning algorithms were employed. We removed superfluous news articles and utilized both random and grid searches to determine the most efficient criteria, leading to improved accuracy and performance. Henceforth, our forthcoming testing will not involve the utilization of imbalanced or default model settings.

By eliminating stop words, we have transformed the statement into one that is clear and free from any ambiguity. The preprocessed text is then transformed into a feature vector, with the verbs being retrieved from the preprocessed text. Finally, the two features are merged. Two types of characteristics have been retrieved. To compute features, using the TF-IDF and Bag-of-Words (Bow) methodologies. The machine learning classifiers used include Support Vector Machine, Random Forest, Extra Tree, Gradient Boosting, XGBoost, Logistic Regression, and Stacking Classifier. The ensemble technique outperforms individual classifiers such as logistic regression and support vector machines. The stacking of random forests, gradient boosting, and extra trees was performed as the basic learners, with logistic regression applied as the final learner.

**CONCLUSION**

Due to limited resources, Urdu lacks a dedicated source for discerning between counterfeit and authentic news. Prior research has predominantly employed smaller datasets, with limited investigation into multi-domain news. This study seeks to address these difficulties by employing a comprehensive and diverse strategy. The key addition of this study is the inclusion of a large dataset consisting of 3284 news articles from nine different domains. This is a substantial improvement compared to previous research that only examined five domains. The data collection process encompassed a range of online sources, including news websites, preexisting articles, and translated news stories from English to Urdu. Thorough verification and meticulous annotation were carried out to guarantee the precision of both genuine and fabricated news. Furthermore, the primary objective is to improve the precision of identifying and verifying false information. The suggested ensemble classifier integrates logistic regression, gradient boosting, random forest, and additional tree techniques. The composite feature vector was created by utilizing verb stacking retrieved from preprocessed text and incorporating TFIDF and BOW approaches.

The efficacy of this method was clearly apparent, especially when incorporating stop words, as evidenced in the conducted studies. The stacked model proved to be the most effective in identifying false information in Urdu, with an accuracy rate of 95.28% and a Matthews correlation coefficient (MCC) of 90.57%, surpassing the performance of other models. Nevertheless, a significant drawback of this work is the lack of utilization of deep learning methodologies. Moreover, the dependence on just three ensemble approaches, namely "bagging", "boosting", and "stacking", suggests the necessity for additional investigation into innovative ensemble methods. This work exclusively focuses on the computation of TF-IDF and BOW features, without exploring alternative approaches such as word embedding, word2vec, or doc2vec. The dataset maintains an equitable 50:50 proportion between fake and real news.

In the future, Urdu misleading information will be detected using deep neural network models. Moreover, our objective is to further improve the dataset to better align with deep learning models. In addition, our proposal includes integrating multiple embeddings such as Fast Text, word embedding, pre-trained embedding, and others. We are interested in utilizing approaches such as ADASYN, SMOTE, or GAN to achieve dataset balance.