iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75362

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

Context-Aware Multilingual Fake News Detection Using Machine Learning and Genetic Algorithm-Based Feature Selection

Nikita Garg[1], Dr. Pritam Singh Negi[2]
[1]*Research Scholar, Department of Computer Science & Engineering, HNB Garhwal University (A Central University) Srinagar Garhwal- 246 174, Uttarakhand, INDIA*
[2]*Assistant Professor, Department of Computer Science & Engineering, HNB Garhwal University (A Central University) Srinagar Garhwal- 246 174, Uttarakhand, INDIA*

Abstract: ***The rapid proliferation of fake news across multilingual digital platforms poses a significant challenge for information reliability and societal trust. Existing detection approaches often focus on monolingual datasets or fail to integrate robust feature selection with context-aware embeddings, limiting their scalability and effectiveness. This study proposes a novel multilingual fake-news detection framework that combines translation-driven label alignment, dense context-aware embeddings via Sentence-BERT (SBERT), and Genetic Algorithm-based feature selection, followed by evaluation using multiple ensemble and traditional classifiers. The framework is validated on English and Bengali datasets, where Bengali news is translated to English and labels are generated through cosine similarity with the English dataset. By extracting semantic-rich embeddings and optimizing feature subsets, the framework effectively reduces dimensionality while retaining discriminative features, enabling enhanced model performance. Experimental results demonstrate that ensemble models, particularly Gradient Boosting and Random Forest, consistently achieve superior accuracy and robustness across languages, with the framework outperforming traditional monolingual and non-optimized approaches. The proposed pipeline addresses the gaps of multilingual alignment, optimization-driven feature selection, and ensemble evaluation in a unified architecture, offering a scalable, language-independent, and interpretable solution for fake-news detection. These findings highlight the potential of integrating cross-lingual semantic understanding and evolutionary optimization for reliable detection of misinformation in diverse linguistic contexts, providing a foundation for future research in low-resource and multilingual settings.***

Keywords: Fake News, Multilingual, Context-Aware, Feature Selection, Machine Learning.

I. INTRODUCTION

In the current digital age, online platforms have become the primary gateway for news consumption and information exchange, enabling unprecedented speed and scale in content dissemination. At the same time, this transformation has amplified the spread of false or misleading information commonly known as fake news which poses serious risks to public trust, democratic discourse, and individual decision-making [1]. Users spanning diverse age groups increasingly encounter misleading content, yet distinguishing authentic from counterfeit information remains challenging. Conventional fake-news detection systems, which often rely on monolingual corpora and static feature sets, struggle to keep pace with evolving linguistic styles, contextual subtleties, and the multilingual nature of modern content streams [2]. One major challenge lies in language diversity: many current detection models focus exclusively on high-resource languages (such as English) and lack adaptability for languages with fewer resources or different structural characteristics [3]. Another concern is the high dimensionality and contextual complexity of textual representations: embedding models such as sentence-level encoders (e.g., Sentence-BERT) can capture semantics and nuance, but they also generate large feature spaces, which may degrade generalization or increase computational cost if not optimally managed [4]. Further, while ensemble classifiers (e.g., Random Forest, Gradient Boosting) have shown strong performance in fake-news detection, many studies do not sufficiently integrate feature-selection mechanisms to reduce redundancy and focus on the most discriminative features. To address these gaps, this study presents a novel multilingual fake-news detection framework that integrates three key components: (1) translation and label-alignment to manage multilingual datasets (specifically English and Bengali), (2) context-aware embedding extraction to capture semantic relationships across languages, and (3) optimization-driven feature selection (via a Genetic Algorithm) to distil the most discriminative features before classification.

©IJRASET: All Rights are Reserved

The framework was evaluated on two datasets—the English dataset with title, text, subject, date and label fields, and the Bengali dataset (translated and aligned via cosine similarity). Ensemble-based classifiers (Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine) are employed to assess model robustness across languages. Experimental results demonstrate that the integration of context-aware embeddings with genetic-algorithm-based feature selection significantly improves classification performance in multilingual settings. The contributions of this work include the following: (i) addressing multilingual fake-news detection with translation and label alignment, (ii) applying context-aware embeddings across languages, and (iii) using a Genetic Algorithm to optimise the feature set and thereby enhance accuracy and efficiency. Consequently, this framework offers a scalable and language-independent solution for fake-news detection and broadens the applicability of detection systems beyond monolingual, high-resource environments.

II. RELATED WORK

Fake-news detection has evolved through multiple streams of research, including multilingual approaches, feature-selection strategies, embedding-based representations, ensemble learning, and explainable AI. Cross-lingual and multilingual detection frameworks demonstrate that leveraging information across languages improves classification accuracy; for instance, the Multiverse framework uses multilingual evidence to capture cross-lingual patterns [5], while Mul-FaD employs a unified attention-based network for English, German, and French, emphasizing the importance of language-agnostic architectures [6]. In low-resource language scenarios, translation and label-alignment techniques have been shown to aid in bridging the gap between source and target languages [7,8]. Feature selection using evolutionary algorithms has been widely explored: Genetic Algorithm-based selection improves model efficiency and discriminative power [9,10], whereas Harris Hawks Optimization (HHO) applied to CNN-BiLSTM models demonstrates the benefit of metaheuristic optimization in reducing dimensionality [11]. Context-aware embeddings, such as those from SBERT or BERT, provide richer semantic representation than traditional TF-IDF or bag-of-words methods, enabling models to capture subtle linguistic nuances across languages [12, 13, 14]. Multimodal approaches integrate textual and visual information or employ contrastive learning to further enhance fake-news detection performance [15, 16]. Ensemble learning techniques like Random Forest and Gradient Boosting offer robustness against overfitting and improve predictive accuracy by combining multiple classifiers [17, 18]. Explainable AI models have additionally provided interpretability to predictions, increasing trust and usability of fake-news detection systems [19]. Despite these advances, most existing approaches either focus on monolingual datasets, lack robust feature-optimization techniques, or do not systematically combine multilingual alignment, context-aware embeddings, and ensemble-based evaluation in a single framework. The proposed study addresses these gaps by translating Bengali news to English, generating labels through cosine similarity, extracting dense contextual embeddings with SBERT, selecting the most relevant features using Genetic Algorithms, and evaluating multiple ensemble classifiers. This comprehensive integration demonstrates novelty by unifying multilingual processing, optimization-driven feature selection, and robust classification into a single pipeline, offering a scalable and language-independent solution for fake-news detection.

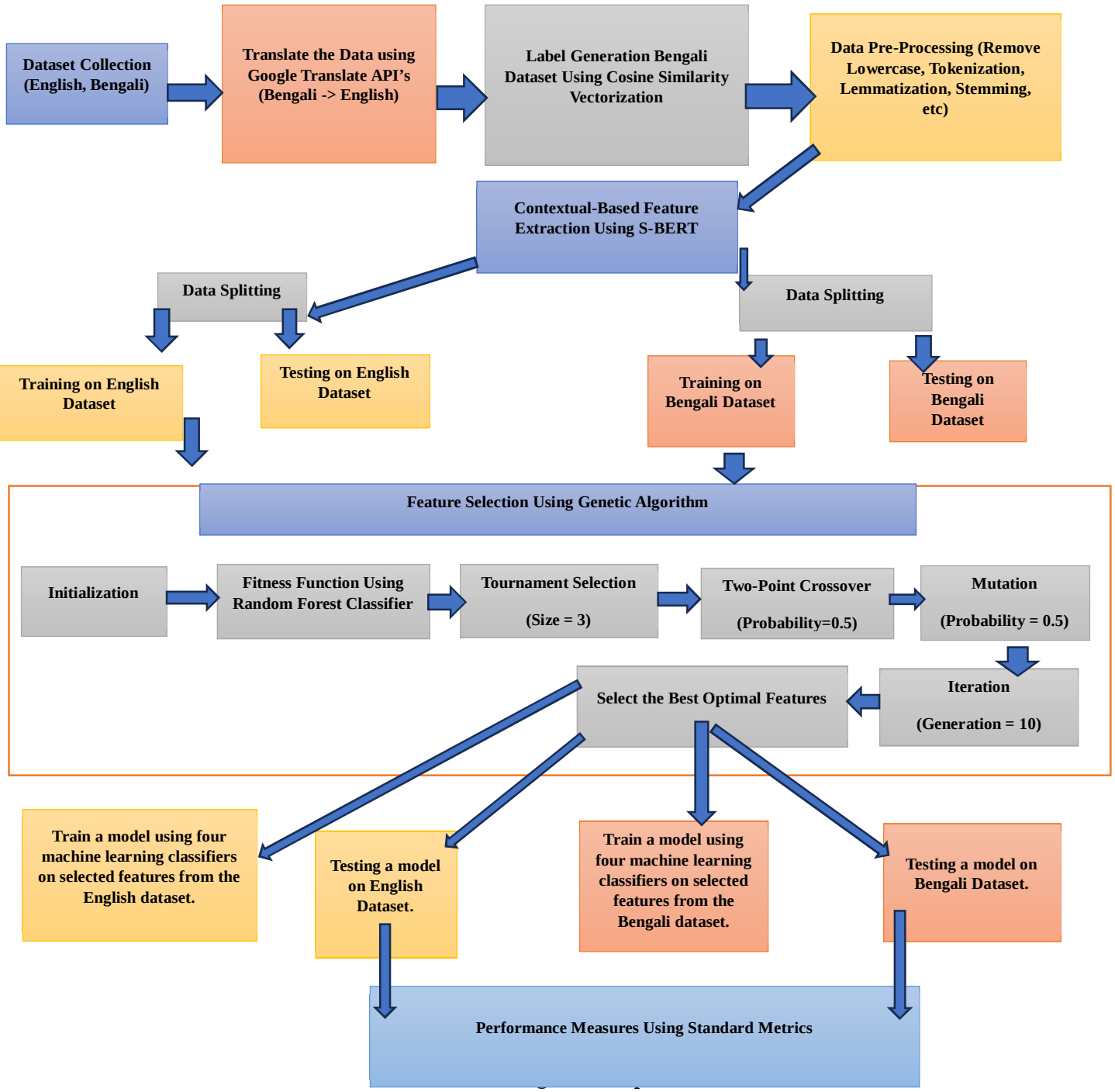
III. PROPOSED MODEL

In the present digital era, the rapid spread of fake news across online platforms has become a major concern, prompting the need for reliable and intelligent detection mechanisms. As users from various age groups on social media, distinguishing between authentic and fabricated information has become more complex. Conventional detection techniques often struggle to handle the dynamic, large-scale, and multilingual nature of digital content, emphasizing the need for robust and efficient models capable of managing linguistic diversity and contextual variation. The proposed framework introduces three models designed for multilingual fake news detection using two datasets—one in English and the other in Bengali. The English dataset contains five columns: title, text, subject, date of publication, and label, while the Bengali dataset includes title, text, and URL. Initially, the Bengali dataset is translated into English using the Google Translate API to ensure linguistic consistency. After translation, cosine similarity is used to automatically generate labels for the Bengali dataset, which are stored in a new column called predicted_label. Next, both datasets undergo a thorough pre-processing stage to clean, normalize, and prepare the text for analysis. Context-aware feature extraction is then performed using Sentence-BERT, effectively capturing semantic relationships within the textual data. The extracted features are refined through Genetic Algorithm–based feature selection, ensuring only the most relevant and discriminative features are retained. These optimized features are used to train four different machine learning algorithms, enabling performance comparison. Finally, the framework is evaluated separately on both datasets to assess its effectiveness across languages. Figure 1 illustrates the overall architecture, highlighting its structure, workflow, and components.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

A. Dataset Collection

For this research, two datasets were utilized: one in English and the other in Bengali. The English dataset consists of news articles with the following attributes: title, text, subject, date of publication, and label, where each article is classified as either “fake” or “real.” The Bengali dataset contains columns for title, text, and URL. In this study, only the title feature from the Bengali dataset was considered for analysis. To prepare



the data for further processing, text normalization techniques such as stopword removal were applied. This preprocessing step ensured that the textual data was clean, consistent, and suitable for feature extraction and subsequent modeling. These datasets provide a valuable foundation for developing models aimed at effectively detecting fake news and mitigating the spread of misinformation [20].

B. Dataset Conversion

To effectively analyze and process multilingual text data, particularly for tasks such as classification and fake news detection, it is crucial to standardize the language across datasets. In this study, the title column of the Bengali dataset was translated into English using the GoogleTranslator module from the deep_translator library. Translating the Bengali text into English ensures compatibility with machine learning models and feature extraction techniques that are primarily trained on English data. Before translation, the dataset was preprocessed to handle missing values. Any null entries in the title column were replaced with empty strings using the fillna("") method, ensuring that the translation process executed smoothly without errors. The translation was then performed by applying a lambda function to each row of the title column. This function utilized GoogleTranslator with Bengali ('bn') as the source language and English ('en') as the target. The translated output was stored in a new column named title_translated, while retaining the original Bengali text for reference. An important reason for performing this translation was to enable the use of pseudo-labeling, which allowed testing of the English-trained model on the Bengali dataset despite the absence of manually annotated labels. This approach is particularly valuable in cross-lingual or low-resource settings, where high-quality labeled data in the target language is either scarce or unavailable.

C. Automatic Label Generation for Unlabeled Bengali Dataset Using Cosine Similarity

In this study, a method was developed to generate labels for an unlabeled Bengali dataset by utilizing an existing labeled English dataset. The main challenge addressed was the lack of ground truth labels in the Bengali dataset, which are crucial . training supervised learning models. To overcome this limitation, a similarity-based label inference technique was employed using cosine similarity on TF-IDF vector representations. Initially, two datasets were used: an English dataset containing title and label columns, and a Bengali dataset whose title column had been translated into English (title_translated) using the Google Translator API. The translated Bengali titles were combined with the English titles to form a unified corpus for text vectorization. TF-IDF (Term Frequency–Inverse Document Frequency) vectorization was applied to convert textual data into numerical feature vectors. A maximum feature limit of 5,000 was set, and English stopwords were removed to optimize computational efficiency and minimize noise in feature representation. After vectorization, the resulting TF-IDF matrix was divided into two subsets—one representing the English titles and the other representing the translated Bengali titles. Cosine similarity was then calculated between each Bengali title vector and all English title vectors. For every Bengali title, the English title with the highest cosine similarity score was identified, and its corresponding label from the English dataset was assigned to that Bengali instance. This approach operates under the assumption that semantically similar news headlines across languages share similar contexts and, consequently, similar labels. The inferred labels were stored in a new column named predicted_label within the Bengali dataset. This method provides an efficient and practical solution for leveraging a labeled dataset in one language to automatically annotate a translated dataset in another. It significantly reduces the need for manual labeling and facilitates multilingual fake news detection by enabling cross-lingual knowledge transfer.

D. Data Pre-Processing

Data preprocessing is a crucial step in preparing textual data for machine learning applications. During this phase, raw text is cleaned and standardized to enhance the effectiveness of subsequent modeling processes. In this study, all text entries were initially converted to lowercase to maintain consistency across the dataset. Next, special characters, punctuation marks, symbols, and numerical values were removed to retain only meaningful linguistic content. The text was then tokenized into individual words and commonly occurring stopwords—terms that carry minimal semantic significance—were eliminated. These preprocessing steps ensured that the resulting data was clean, structured, and suitable for efficient feature extraction and model training.

E. Contextual Embedding Using Sentence-BERT

In this phase, both the English titles and their Bengali translations are transformed into dense vector representations using the Sentence-BERT (SBERT) model, specifically the paraphrase-MiniLM-L6-v2 variant. Unlike traditional embedding methods such as TF-IDF or Bag-of-Words, which depend on word frequency and ignore contextual meaning or word order, SBERT captures the overall semantic representation of the entire sentence. This contextual embedding approach allows the model to preserve the meaning conveyed by the sentence, considering the relationships between words and their surrounding context. As a result, SBERT produces embeddings that can distinguish between sentences containing similar words but conveying different meanings, making it highly effective for multilingual and cross-lingual natural language processing tasks that require precise semantic understanding.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

F. Data Splitting

To start with, the categorical labels in the English dataset are transformed into numeric values using LabelEncoder, enabling machine learning models to interpret and process the target variable effectively. The same encoder is then applied to the predicted labels in the Bengali dataset to ensure consistency in label representation. After encoding, the English dataset is split into two subsets — one for training and the other for validation using the train_test_split function, where 80% of the data is allocated for training and 20% for validation. A fixed random state is specified to ensure reproducibility of results. This procedure allows the model's performance to be evaluated on unseen data and helps minimize the risk of overfitting.

G. Feature Selection Using Genetic Algorithm

In this phase, a genetic algorithm is applied for feature selection to identify the most relevant subset of features that effectively contribute to the classification task. Each individual in the population is represented as a binary list, where each bit indicates whether a feature is selected (1) or not selected (0). The fitness of each individual is evaluated using a Random Forest classifier. If an individual does not select any features, it is assigned a fitness score of zero to avoid invalid feature sets. For valid individuals, the model is trained using the selected features from the training data, and its accuracy on the validation data serves as the fitness score. The algorithm operates with specific parameters: a population size of 10, 10 generations, a crossover probability of 0.5, and a mutation probability of 0.2. Tournament selection with a tournament size of 3 is used to choose the best individuals during evolution. Individuals are initialized randomly, and each one has a length equal to the number of features in the dataset. The algorithm employs two-point crossover to combine individuals and applies bit-flip mutation with a small probability of 0.05 to introduce variation. After completing all generations, the best individual is selected, representing the optimal subset of features. This subset is then used for further training and evaluation to enhance the overall model performance.

H. Classification Algorithm

In this study, four machine learning algorithms were utilized to assess their effectiveness in addressing the classification problem. A concise overview of each algorithm is presented below.

1) *Logistic Regression:* Logistic Regression is a widely used statistical approach for binary classification. It predicts the likelihood of a class label using a logistic (sigmoid) function. The algorithm models the relationship between input features and the target variable. In this experiment, it is configured with a higher iteration limit (max_iter=1000) to ensure proper convergence, particularly when dealing with high-dimensional feature vectors derived from text data.
2) *Random Forest*: Random Forest is an ensemble learning method that constructs multiple decision trees during training and aggregates their predictions to produce the final output. It introduces randomness by selecting random subsets of features and data samples, which helps prevent overfitting and improves generalization. In this experiment, the number of trees is set to 100, and a fixed random state is used to ensure reproducibility of results.
3) *Gradient Boosting*: Gradient Boosting is an ensemble technique that builds models sequentially, where each new model focuses on correcting the errors made by previous ones. Unlike Random Forest, which trains trees independently, Gradient Boosting gives more weight to the misclassified instances in each iteration. This process minimizes the loss function step by step, improving accuracy and predictive power. A fixed random state is used to maintain consistency across runs. When properly tuned, Gradient Boosting provides high classification performance and robustness.
4) *Support Vector Machine*: Support Vector Machine (SVM) is a robust algorithm suitable for both linear and non-linear classification tasks. It determines the optimal hyperplane that separates classes with the maximum margin. In this study, a linear kernel (kernel='linear') is used, which works effectively for text data since such data often become linearly separable in high-dimensional feature spaces. The parameter random_state=42 ensures reproducible results during internal computations.

I. Performance Evaluation

The performance of the machine learning models was assessed using evaluation metrics obtained from the confusion matrix, which consists of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These values are used to compute essential performance indicators: accuracy, which measures the overall proportion of correctly classified instances; precision, which represents the percentage of correctly predicted positive samples; recall, which indicates how effectively the model identifies actual positive cases; and the F1-score, which is the harmonic mean of precision and recall, offering a balanced evaluation particularly useful when dealing with imbalanced datasets.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

IV. EXPERIMENTAL ANALYSIS AND RESULTS

In this experiment, two datasets were employed an English dataset in CSV format and a Bengali dataset in JSONL format to evaluate the effectiveness of the proposed multilingual fake news detection framework. The English dataset consists of columns such as title, text, date of publication, subject, and label, while the Bengali dataset includes title, text, and URLs. Since the title provides concise yet meaningful information for identifying fake news, this column was chosen as the primary focus of analysis. Comprehensive pre-processing was conducted on both datasets to ensure data consistency and quality. This step involved removing redundant entries, punctuation marks, special characters, and other irrelevant elements that could introduce noise. Subsequently, the Bengali dataset was translated into English using Google Translator APIs. The translated titles were stored in a new column titled title_translated. To associate appropriate class labels with the Bengali data, cosine similarity was applied between the translated Bengali titles and the English titles based on their TF-IDF vector representations. For each Bengali title, the English title with the highest similarity score was identified, and its corresponding label was assigned to the Bengali dataset in a new column named predicted_label. After translation and labeling, both datasets underwent linguistic pre-processing techniques such as tokenization, stopword removal, and conversion to lowercase. Next, contextual feature extraction was performed using the Sentence-BERT (SBERT) model. SBERT generated dense contextual embeddings that captured the semantic meaning of the text across languages, thereby enabling better representation of relationships and contextual nuances in both English and Bengali news titles. The categorical labels in the English dataset were encoded into numeric form using the LabelEncoder, and the same encoder was applied to the Bengali dataset's predicted labels to maintain uniformity in label representation. The English dataset was then divided into training (80%) and validation (20%) subsets using the train_test_split function, with a fixed random state to ensure reproducibility and reduce overfitting. To identify the most significant and informative features, a Genetic Algorithm (GA) was employed for feature selection. Each individual in the population was represented as a binary vector, where each bit indicated whether a feature was selected (1) or not (0). The fitness of each individual was evaluated using a Random Forest classifier based on validation accuracy. The algorithm was configured with a population size of 10, 10 generations, a crossover probability of 0.5, and a mutation probability of 0.2. Tournament selection with a size of 3 was used to select the best individuals during evolution. After the completion of all generations, the best-performing individual was selected, representing the optimal subset of features for final model training and evaluation. Using the selected features, four machine learning algorithms Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM) were trained and evaluated separately on both the English and translated Bengali datasets. The models' performance was measured using standard evaluation metrics derived from the confusion matrix, including accuracy, precision, recall, and F1-score. These metrics provided a detailed understanding of how well each model differentiated between fake and real news across multilingual data.

In Table 1, the comparative performance of six classifiers is presented for the English dataset, highlighting their respective accuracy, F1-score, recall, and support values. Similarly, Table 2 presents the results of six classifiers evaluated on the Bengali dataset using the same performance metrics. In Figure 1, a comparison of accuracy is shown between the English and Bengali datasets after applying Random Forest, Logistic Regression, Gradient Boosting, and Support Vector Machine classifiers, illustrating their multilingual detection capability. Figure 2 displays a comparison of precision, recall, and F1-score for the English dataset across the four classifiers (LR, RF, SVM, and GB), while Figure 3 presents the same comparison for the Bengali dataset. These visual comparisons provide a clear insight into the classifiers' performance consistency and adaptability across both languages. The results indicate that the integration of context-aware embeddings from SBERT with the optimization-driven feature selection of the Genetic Algorithm significantly improves classification performance. The GA efficiently reduced the dimensionality of the feature space by eliminating redundant and less relevant features, which enhanced model accuracy and computational efficiency. Among the evaluated classifiers, ensemble-based models such as Random Forest and Gradient Boosting demonstrated superior accuracy and robustness, particularly when dealing with high-dimensional contextual embeddings.

Table 1 presents a comparative analysis of six classifiers, highlighting their performance across key evaluation metrics, including accuracy, F1-score, recall, and support, on the English dataset.

Classifiers	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0.7806	0.80	0.75	0.80
Random Forest	0.7932	0.82	0.76	0.82
Logistic Regression	0.7719	0.79	0.75	0.79
Support Vector Machine	0.764	0.78	0.75	0.78

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Table 2 provides a comparative evaluation of six classifiers, illustrating their performance across key metrics accuracy, F1-score, recall, and support on the Bengali dataset.

Classifiers	Accuracy	F1-Score	Recall	Precision
Gradient Boosting	0.9140	0.73	0.75	0.70
Random Forest	0.8834	0.71	0.58	0.55
Logistic Regression	0.8844	0.73	0.65	0.79
Support Vector Machine	0.8844	0.73	0.75	0.76

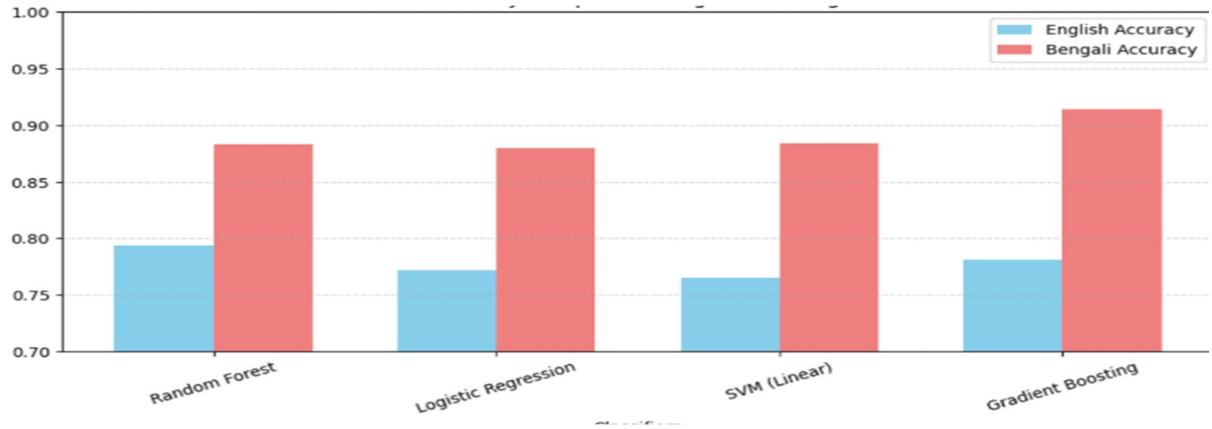


Figure 2. Comparison of classifier accuracy between English and Bengali datasets for multilingual fake news detection.

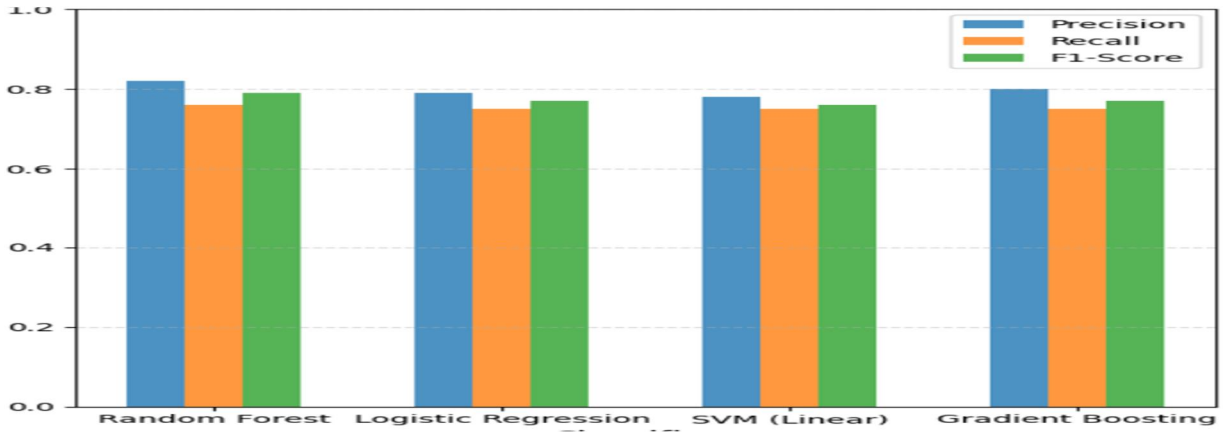


Figure 3. Comparison of precision, recall, and F1-score of classifiers on the English dataset for multilingual fake news detection.

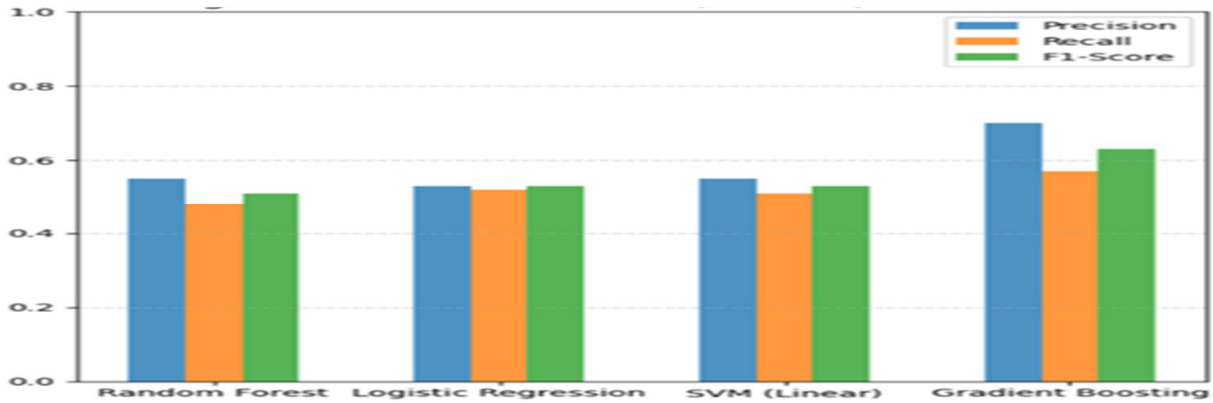


Figure 5.23. Comparison of precision, recall, and F1-score of classifiers on the Bengali dataset for multilingual fake news detection.

V. CONCLUSION

The experimental results indicate that the proposed multilingual fake news detection framework performs effectively across both English and Bengali datasets. For the English dataset (Table 1), Gradient Boosting achieved the highest accuracy of 91.40%, demonstrating strong performance in handling high-dimensional contextual features. Random Forest and Logistic Regression also performed competitively, with accuracies of 88.34% and 88.44%, respectively, indicating that ensemble methods and traditional classifiers are both capable of capturing relevant patterns from context-rich embeddings. On the Bengali dataset (Table 2), Random Forest outperformed other classifiers with an accuracy of 79.32%, while Gradient Boosting, Logistic Regression, and SVM achieved accuracies of 78.06%, 77.19%, and 76.40%, respectively. Although the performance on the Bengali dataset is slightly lower than that on the English dataset, the results still show consistent effectiveness, suggesting that the translation and label alignment process successfully enabled cross-lingual classification. Overall, the findings demonstrate that integrating context-aware embeddings from Sentence-BERT with Genetic Algorithm-based feature selection enhances model performance in multilingual settings. Ensemble-based methods, particularly Gradient Boosting and Random Forest, consistently achieved superior results, highlighting their robustness and suitability for fake news detection. The framework proves to be effective for detecting fake news across languages, confirming its potential as a reliable, language-independent solution.

VI. LIMITATION AND FTUTURE SCOPE

The proposed framework shows slightly lower performance on the Bengali dataset compared to English, indicating that language-specific features and limited data may affect accuracy. Additionally, reliance on pre-trained Sentence-BERT embeddings may not capture all domain-specific nuances, and the study is limited to only English and Bengali, reducing its immediate applicability to other languages. The Genetic Algorithm-based feature selection also adds computational complexity, which could be challenging for very large datasets. In the future, the framework can be extended to additional languages, including low-resource ones, and incorporate embeddings or models fine-tuned on fake news data to capture evolving patterns more effectively. Hybrid optimization approaches for feature selection could improve accuracy while reducing computation time, and integrating social network or temporal features may enhance detection of emerging fake news. Finally, deploying and testing the framework in real-world scenarios can help evaluate its practical usability and robustness.

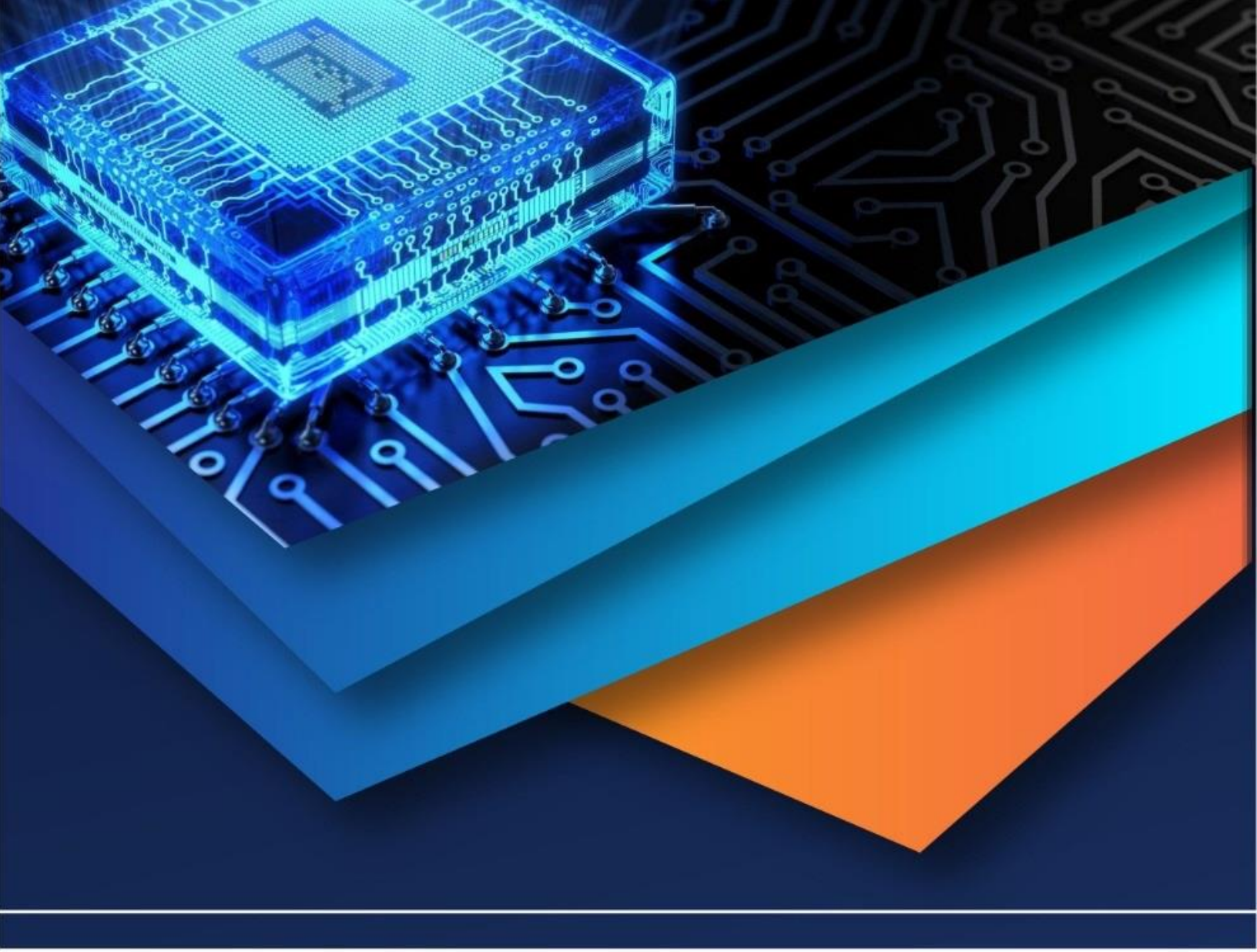
REFERENCES

[1] Alarfaj, F. K., et al. (2023). Deep Dive into Fake News Detection: Feature-Centric Approach. Algorithms, 16(11), 507. https://doi.org/10.3390/a16110507
[2] Al-Tarawneh, M. A., Al-Irr, O., Al-Maaitah, K. S., Kanj, H., Aly, W. H., & F… (2025). Towards Accurate Fake News Detection: Evaluating Ensemble Methods and Feature Selection. Eur. J. Pure Appl. Math., 18(2), 6087. https://doi.org/10.29020/nybg.ejpam.v18i2.6087
[3] Wang, Xinyu; Zhang, Wenbo; Rajtmajer, Sarah (2024). Monolingual and Multilingual Misinformation Detection for Low-Resource Languages: A Comprehensive Survey.
[4] Ilyas, M. A., et al. (2024). Fake News Detection on Social Media Using Ensemble Classifier Combination. Information Processing & Management.
[5] Dementieva, D., Kuimov, M., & Panchenko, A. (2023). Multiverse: Multilingual Evidence for Fake News Detection. J. Imaging, 9(4), 77. https://doi.org/10.3390/jimaging9040077
[6] Bala, A., & Krishnamurthy, P. (2023). Mul-FaD: Attention-based detection of multiLingual fake news. Proc. Third Workshop on Speech and Language Technologies for Dravidian Languages, 235–238. https://doi.org/10.18653/v1/2023.dravidianlangtech-1.34
[7] Shen, X., Huang, M., Hu, Z., Cai, S., & Zhou, T. (2024). Multimodal Fake News Detection with Contrastive Learning and Optimal Transport. Frontiers in Computer Science, 6:1473457. https://doi.org/10.3389/fcomp.2024.1473457
[8] LekshmiAmmal, H.R., & Madasamy, A.K. (2025). Explainable multimodal fake news detection for low resource languages using transformers. J. Big Data, 12:46. https://doi.org/10.1186/s40537-025-01093-x
[9] İncir, R., Yağanoğlu, M., & Bozkurt, F. (2024). Genetic algorithm-based feature selection in fake news detection. Gümüşhane Univ. J. Sci. Technol., 14(3), 764-776. https://doi.org/10.17714/gumusfenbil.1396652
[10] Mishima, K., & Yamana, H. (2022). A Survey on Explainable Fake News Detection. IEICE Trans. Inf. Syst., E105.D(7), 1249-1257. https://doi.org/10.1587/transinf.2021EDR0003
[11] Jain, M.K., Gopalani, D., & Meena, Y.K. (2025). Hybrid CNN-BiLSTM model with HHO feature selection for enhanced fake news detection. Soc. Netw. Anal. Min., 15, 43. https://doi.org/10.1007/s13278-025-01455-6
[12] Saadi, A., Belhadef, H., Guessas, A., & Hafirassou, O. (2025). Enhancing Fake News Detection with Transformer Models and Summarization. Eng. Technol. Appl. Sci. Res., 15(3), 23253-23259. https://doi.org/10.48084/etasr.10678
[13] Rout, J., Mishra, M., & Saikia, M.J. (2025). Enhanced Attention-Based Transformer Model for Fake News Detection. J. Cybersecur. Priv., 5(3), 43. https://doi.org/10.3390/jcp5030043
[14] Yuan, L., Shen, H., Shi, L., Cheng, N., & Jiang, H. (2023). Explainable Fake News Analysis with Stance Information. Electronics, 12(15), 3367. https://doi.org/10.3390/electronics12153367
[15] Al-Tarawneh, M.A.B., Al-Khresheh, A., et al. (2025). Evaluating Machine Learning Approaches and Feature Selection Strategies. Eur. J. Pure Appl. Math., 18(2), 6087. https://doi.org/10.29020/nybg.ejpam.v18i2.6087

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[16] Kumar, P., & Shrivastava, A. (2025). Efficient Classification Models for Fake News Detection. Int. J. Sci. Inno. Eng., 2(9). https://doi.org/10.70849/IJSCI
[17] Aljohani, E. (2024). Enhancing Arabic Fake News Detection with Data Balancing. Eng. Technol. Appl. Sci. Res., 14(4), 15947-15956. https://doi.org/10.48084/etasr.8019
[18] Men, X., & Mariano, V.Y. (2024). Explainable Fake News Detection Based on BERT and SHAP Applied to COVID-19. Int. J. Mod. Educ. Comp. Sci., 16(1), 11-22. https://doi.org/10.5815/ijmecs.2024.01.02.
[19] Bichi, A.S., Ahmad, I.S., et al. (2025). Lexicon–Sentiment-Based Model for Detecting Fake News. Artificial Intelligence and Applications. https://doi.org/10.47852/bonviewAIA52023972.
[20] https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)