I. J. Engineering and Manufacturing, **2026, 4, 97-105**
Published Online on August 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2026.04.07

Multilingual Fake News Detection Using Machine Learning with Contextual-Based Feature Extraction

Nikita Garg*
Department of Computer Science & Engineering, HNB Garhwal University (A Central University), Srinagar Garhwal-246 174, Uttarakhand, India
E-mail: nikita.garg2706@gmail.com
ORCID iD: https://orcid.org/0009-0009-1418-2611
*Corresponding Author

Pritam Singh Negi
Department of Computer Science & Engineering, HNB Garhwal University (A Central University), Srinagar Garhwal-246 174, Uttarakhand, India
E-mail: negipritam@hnbgu.ac.in
ORCID iD: https://orcid.org/0009-0004-5703-6212

Received: March 5, 2026; Revised: April 12, 2026; Accepted: July 16, 2026; Published: August 8, 2026

Abstract: Fake news has become a major challenge in today's digital environment, particularly in languages where labeled data is limited. Most existing research has primarily focused on English due to the easy availability of annotated datasets, whereas low-resource languages such as Bengali remain underexplored. This study presents a multilingual approach for fake news detection using machine learning with contextual-based feature extraction. The proposed method integrates n-gram techniques with sentence-level contextual embeddings to capture both word-level patterns and semantic meaning. Since labeled data is not available for the Bengali dataset, a translation-based strategy is employed, followed by a pseudo-labeling process to assign labels automatically. The models are trained on English news titles and subsequently evaluated on both English and Bengali datasets to examine their cross-lingual effectiveness. The experimental findings indicate that ensemble-based classifiers such as Random Forest and Gradient Boosting achieve reliable performance across both languages. In some cases, the results for Bengali data are comparable or slightly better than those for English. The study demonstrates that effective fake news detection is possible in low-resource languages using short text data without relying on manually labeled datasets. The proposed approach provides a simple and efficient solution for multilingual fake news detection in data-scarce environments.

Index Terms: Fake News Detection, Multilingual Natural Language Processing, Feature Extraction, Contextual Embeddings, N-gram, Machine Learning

1. Introduction

In today's digital environment, the rapid spread of misinformation across online platforms has increased the need for effective fake news detection systems. Studies have shown that false information spreads faster and reaches a wider audience than true information, highlighting the seriousness of the issue[1]. In addition, several surveys have emphasized the importance of automated fake news detection using ML techniques [2].

This challenge becomes more significant in low-resource languages such as Bengali, where labelled datasets and linguistic resources are limited. Semi-supervised approaches, including pseudo-labeling have been explored to address the lack of annotated data by utilizing unlabeled samples [3]. Furthermore, cross-lingual representation learning and multilingual embeddings enable knowledge transfer across languages, helping to reduce the gap between high-resource and low-resource settings [4].

Low-resource languages remain underrepresented in NLP due to the limited availability of annotated corpora and linguistic tools, which creates challenges for model development [5]. FND has evolved from traditional approaches to more advanced ML techniques, improving classification performance in recent years [6]. More recently, transformer-

based architecture has further enhanced the ability to capture contextual relationships in textual data [7].

To overcome language-related limitations, recent research has explored low-resource NLP techniques. Such as transfer learning & data augmentation [8]. Multilingual models such as BERT have demonstrated the capability to learn shared representations across multiple languages, supporting cross-lingual learning tasks [9]. In addition, cross-lingual learning approaches, including transition-based methods, have shown effectiveness in transferring knowledge from high-resource to low-resource languages [10].

A significant research gap exists in developing lightweight and scalable fake news detection systems for low-resource languages. While state-of-the-art Transformer models (e.g., mBERT, XLM-R) provide high performance, they are computationally expensive and require large, annotated datasets. The novelty of this study lies in introducing a resource-efficient pipeline that fuses SBERT and N-gram features specifically for short-text news titles [15,16]. This approach bypasses the need for native Bengali labels through a translation-based framework, a strategy increasingly seen as effective in recent survey literature [17]. By combining lexical word-group patterns with deep semantic sentence embeddings, the proposed framework offers a practical and low-compute solution for real-time misinformation analysis in data-scarce environments.

Despite these advancements, most existing studies focus on high-resource or moderately resourced language survey on text classification highlight the need for more adaptable approaches across diverse linguistics settings [13]. Recent research also emphasizes the growing importance of multilingual FND, particularly in underrepresented languages [14].

This paper is organised into six main sections. Section 2 presents the related work around FN and multilingual approaches. Section 3 describes the methodology, including the dataset, feature extraction, and modeling approach. Section 4 reports the results and provides a discussion of the findings. Section 5 concludes the study by summarizing the main contributions. Finally, section 6 outlines the limitations of the current work and suggests directions for future work.

2. Related Work

The use of digital platforms has increased a lot in recent years, and information spreads very quickly. This also includes false news. Because of this FND has become important. But at the same time, misleading and false news has also increased. In many cases, false information reaches more people faster than true news, which makes this problem serious [1]. Due to this reason, FND has become an important research area.

Earlier, most of the work in this field was based on traditional ML methods. These methods worked, but they were not always enough for complex text data. So, researchers started using some advanced techniques. One of them is semi-supervised learning. In this, pseudo-labelling is used so that unlabeled data can also be used along with labelled data [3]. Also, multilingual and cross-lingual approaches are used, where models can learn from more than one language [4].

Still, low-resource languages face many problems. The main issue is that there are very limited labelled data and fewer linguistic resources available [5]. Because of this, models do not perform as well as they do in high-resource languages. Deep learning and transformer-based models have shown good results, but mostly for languages where enough data is available [6].

The use of digital platforms has increased a lot in recent years, and information spreads very quickly, also includes false news. Because of this, FND has become important. To improve this, researchers have tried transfer learning. In this approach, knowledge from high-resource languages is transferred to low-resource languages [8]. Multilingual models like BERT are also helpful, as they can work with multiple languages & capture patterns across support knowledge sharing between languages and improve performance [10].

Apart from this, ensemble methods are also used. These methods combine multiple models to improve accuracy [11]. Pseudo-labelling methods have also improved over time and are now effective in semi-supervised learning [12]. Recently, context-aware and multilingual approaches combined with feature selection techniques such as genetic algorithms have also been explored for improving FND performance across languages.

Existing literature often focuses on full-length articles in high-resource languages; however, transferring these techniques to Bengali titles remains challenging due to semantic drift. Modern multilingual models like mBERT and XLM-R have emerged as powerful tools, but their performance often declines for languages with complex syntax and limited morphological resources. Our framework addresses these limitations by providing a scalable alternative that captures both lexical patterns and semantic intent without the high-computer requirements typically associated with Transformer-based architecture. This comparison highlights that a hybrid pipeline can achieve competitive results on short-text titles while maintaining resource efficiency. However, most research still focuses on high-resource languages, while low-resource languages receive less attention.

3. Methodology

In this work, FND is carried out using data from English and Bengali. The Bengali data does not have labels, so it is first converted into English. After that, basic cleaning is done to make the text ready for use to get useful information

from the text. N-grams & contextual embeddings are used. N-grams help in finding common word patterns, while embeddings help in understanding the meaning of the text. Next, pseudo-labelling is used. In this step, the model gives labels to the Bengali data based on its own predictions. This helps in using the data even without manual labels. The model is trained on the English data & then checked on both English and Bengali datasets to see how well it performs on both languages. The full process is shown in Fig. 1, Proposed Framework for Multilingual Fake News Detection.

3.1. Dataset Collection

Two datasets are used; these are taken from Kaggle. One is an English dataset denoted as D_{en} and one is the Bengali dataset, denoted as D_{bn}. In the English dataset $D_{en} = \{(x_i^{en}, y_i)\}_{i=1}^{N}$ contains news titles with labels, which are already given 0 means fake and 1 means real. But the Bengali dataset $D_{bn} = \{x_j^{bn}\}_{j=1}^{M}$ includes titles, text, and summaries but does not have any labels. It only contains title, text and summary. To make the process simple, only the title part is used from both datasets. Before using the data, some basic cleaning is done. All text is changed into lowercase, extra spaces are removed, and unwanted symbols are also deleted. Common stopwords are also removed so that the text becomes clearer. The English dataset is used to train the model because it has proper labels. For the Bengali dataset, translation is done first, and then pseudo-labelling is used so that it can also be used in the process. The English dataset consists of approximately 5,000 news titles, while the Bengali dataset contains 3,000 titles sourced from diverse domains including politics, health, and social media. Both datasets maintain a near-equal class balance between 'fake' and 'real' instances to prevent model bias. Furthermore, the use of only news titles is justified as they provide concise and impactful information, are readily available in real-time scenarios such as social media platforms and help reduce computational complexity while still capturing essential cues for effective fake news detection.

3.2. Data Conversion

In this part, Bengali titles are changed into English, this is done so that both data can be used in the same format. For this, a translator tool is used. Some titles were empty, so they were filled with blank text to avoid errors. Then each Bengali title is translated one by one. The translated text is saved in a new column. The original Bengali text is also kept. After this, all titles are in English, so it becomes easy to use the same model. The model trained on English data is then used to give labels to the Bengali data. This step is useful because Bengali data does not have labels. To address the risk of semantic drift, where the original meaning might change during translation SBERT is utilized. SBERT focuses on sentence-level intent rather than word-for-word translation, ensuring that the semantic integrity of Bengali news is preserved when transformed into the English feature space [16].

3.3. Data Pre-Processing

Here, the text is cleaned, and all words are written in small letters. After that, punctuation, symbols and numbers are removed. Then the text is broken into small parts (words). Common words like “is”, “the”, and “and” are also removed because they are not very useful. After this, the text becomes simple & ready to use for further steps.

3.4. Context-Based Feature Extraction

In text work, the data is in words, but the model needs numbers. So, the text is changed into numbers before using it. In this study, two methods are used together. One method looks at small word groups, and other one helps to understand the meaning of the text. Each sentence or document is then changed into a set of values. These values are later used by the model for learning. Let the dataset be denoted as $D = \{d_1, d_2, \dots, d_n\}$, where each document $d_i$consists of a token sequence $T_i = [w_1, w_2, \dots, w_m]$. The objective is to transform each document into a feature vector $F_i \in \mathbb{R}^d$suitable for machine learning models.

3.5. N-Gram Vectorization

An n-gram looks at small groups of words that come together in a sentence. These word groups help in seeing common patterns in the text. An n-gram represents a sequence of *n* consecutive words, such as unigrams, bigrams, and trigrams, where expressions like “fake news” form a bigram that conveys a specific concept. For a document $d_i$with m tokens, an n-gram is defined as $\{(w_j, w_{j+1}, \dots, w_{j+n-1}) \mid 1 \le j \le m - n + 1\}$, where $w_j$denotes the j-th token. After extracting all n-grams, they are transformed into numerical vectors using CountVectorizer, which builds a sparse matrix based on word frequencies. To keep the feature space manageable while retaining key patterns, only the top $k = 3000$ The most frequent n-grams are selected. The resulting vector $v_i = \text{CountVectorizer}(\text{n-gram}_i)$captures essential phrase-level cues and local dependencies useful for downstream tasks such as fake news detection.

3.6. Contextual Embeddings Using Sentence Transformers

Traditional embeddings such as Word2Vec and GloVe assign the same vector to a word no matter where it appears, making them unable to reflect changes in meaning across different contexts. To overcome this limitation, this study uses contextual embeddings that adapt a word’s representation based on the sentence in which it occurs. Sentence-BERT (SBERT) is used to generate these context-aware sentence vectors, producing a single dense embedding that captures

how words relate to one another within the full sentence. For each document d_i, SBERT computes a sentence embedding $s_i = f_{\text{SBERT}}(d_i)$, where $s_i \in \mathbb{R}^d$. These embeddings offer a richer semantic understanding by encoding both meaning and context, making them more effective for tasks like multilingual fake news detection.

3.7. Combined Feature Representation

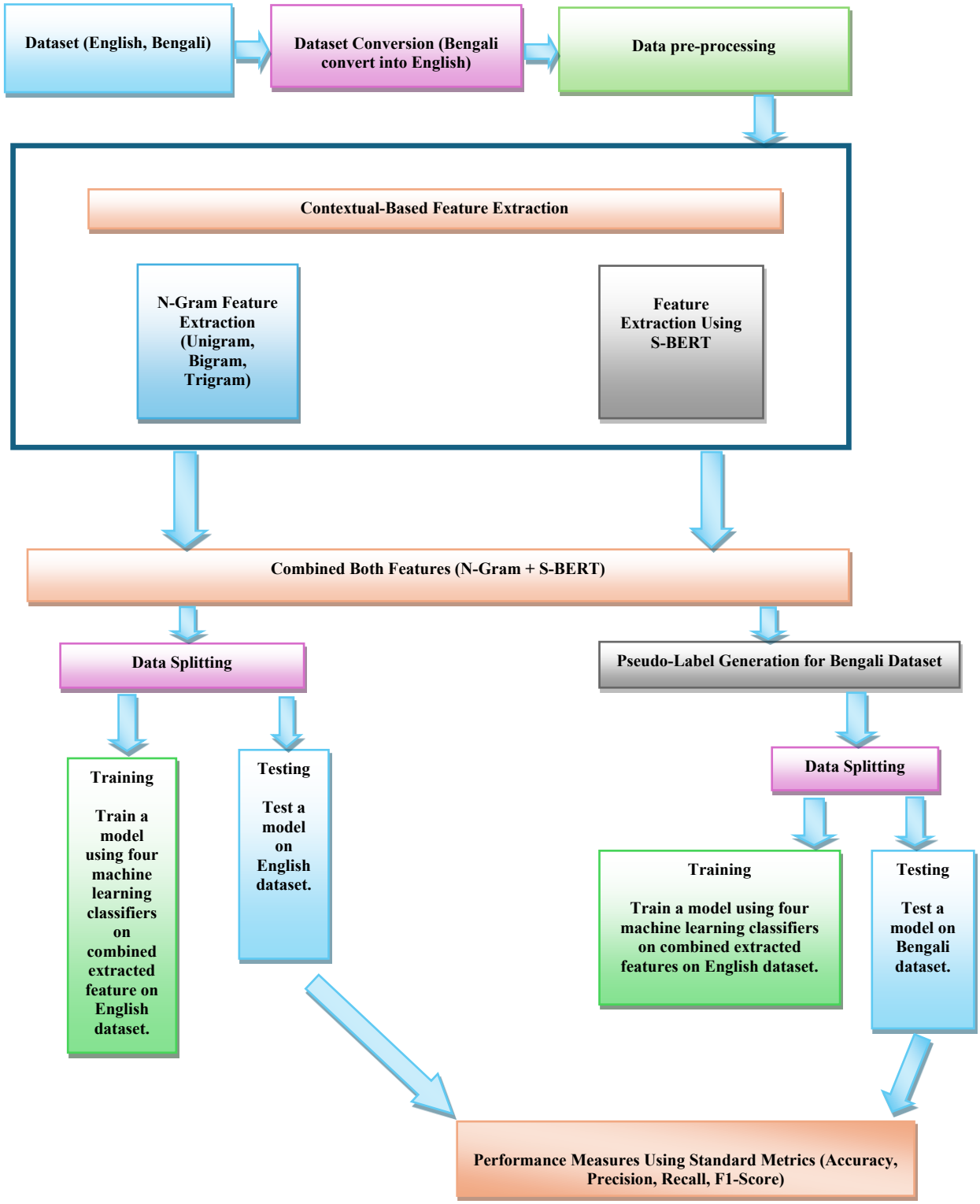

Fig. 1. Proposed Framework for Multilingual Fake News Detection

To integrate both surface-level patterns and deeper semantic cues, the final feature vector for each document is formed by concatenating the n-gram vector $v_i$with the contextual embedding s_i, expressed as $F_i = [v_i \mid s_i]$, where $F_i \in$

$\mathbb{R}^{k+d}$. Here, $k$denotes the number of selected n-grams and $d$represents the dimensionality of the SBERT embeddings**.** The final feature vector for each document is the result of fusing lexical word patterns and semantic embeddings. The combined feature vector is defined in (1):

$$F_i = V_i \oplus S_i \tag{1}$$

where Vi represents the N-gram feature vector, Si represents the SBERT embedding, and $\oplus$ denotes the vector concatenation operator.[17].

This fusion allows the model to draw on complementary strengths: n-grams highlight recurring word sequences and syntactic structure, while SBERT captures the broader semantic relationships within the sentence. The combined features give a better idea of the text. It helps the model understand news from different languages more clearly.

3.8. Data Splitting

Data is divided into two parts; one part is used to train the model. The other part is used to test the training data, which helps the model learn, and the testing data is used to verify the results. A fixed random state of 42 was used to ensure the split was consistent and reproducible across all runs. Beyond the initial 80-20 split, a 5-fold cross-validation strategy was implemented during the training phase. This ensures the stability of the results across multiple runs and confirms that the model's performance is not dependent on a specific data partition.

3.9. Classification Algorithms

To evaluate the effectiveness of the proposed multilingual fake news detection framework, four supervised machine learning algorithms were employed: logistic regression, random forest, gradient boosting, and support vector machine. The English dataset, containing manually annotated labels, was used as the primary training data. Since the Bengali dataset lacked ground-truth labels, it was first translated into English and then subjected to text preprocessing, including data cleaning and normalization. Linguistic information was represented using a combination of N-gram features and contextual embeddings to capture both local word patterns and semantic relationships within the news articles. A pseudo-labelling strategy was subsequently applied to assign labels to the unlabeled Bengali samples based on model predictions. The resulting pseudo-labelled data were incorporated into the training process, allowing the classifiers to learn from both manually labelled instances. All classification models were trained and evaluated under the same experimental settings to provide a consistent comparison of their performance for multilingual fake news detection.

3.9.1. Logistic Regression

Logistic Regression is a supervised classification algorithm that predicts the probability of an input sample belonging to one of two classes. In this study, the classifier was trained using the labelled English news articles after feature extraction through N-gram and contextual embeddings. After generating pseudo-labels for the translated Bengali news articles, these samples were also utilised to improve the learning process. Logistic Regression established a linear decision boundary based on the extracted textual features, making it suitable for binary classification problems, such as identifying fake and real news. The model was selected due to its computational efficiency and its ability to deliver reliable performance on high-dimensional textual data.

3.9.2. Random Forest

Random Forest is an ensemble learning algorithm that combines the predictions of multiple decision trees to produce the final classification result. Each decision tree is constructed using different subsets of the training data and feature space, enabling the model to capture diverse characteristics of textual information. In the proposed framework, a random forest was trained using the linguistic features obtained from N-grams and contextual embeddings extracted from the English dataset, together with the pseudo-labelled Bengali samples. The aggregation of multiple trees improves classification stability and reduces the likelihood of overfitting, making the model suitable for multilingual fake news detection.

3.9.3. Gradient Boost

Gradient Boosting is an ensemble technique that sequentially constructs decision trees, where each new tree focuses on correcting the prediction errors of the previously generated trees. This iterative learning process enables the model to improve its classification capability over successive stages. In the proposed methodology, gradient boosting utilized the feature representations generated from N-grams and contextual embeddings after preprocessing and pseudo-label generation. By continuously minimizing classification errors, the models effectively and accurately classify genuine news articles across both English and translated Bengali datasets.

3.9.4. Support Vector Machine

Support Vector Machine is a supervised learning algorithm that identifies an optimal separating hyperplane to distinguish different classes. While maximizing the margin between them, owing to its effectiveness in handling high-dimensional textual representations, the support vector machine is widely adopted for text classification tasks. In this work, the classifier was trained using the combined feature representation obtained from N-gram and contextual embeddings extracted from the labelled English dataset and the pseudo-labelled Bengali news articles. The trained model was subsequently evaluated on both datasets to examine its capability to accurately classify multilingual fake and real news, while maintaining robust generalization performance.

3.10. Pseudo Label Generation

A significant challenge in processing low-resource Bengali news is the lack of gold-standard human-annotated labels. To address this, a pseudo-labelling strategy was adopted. After preprocessing the Bengali dataset to align with the English feature space, the optimised Logistic Regression model, which demonstrated the highest performance on the source English data, was used as a labelling engine. This model generates silver-standard labels by predicting the probability of each Bengali instance being 'fake' or 'real'. While this approach enables cross-lingual transfer, we acknowledge the inherent risk of error propagation, where any initial classification biases from the source model may be transferred to the target dataset. To minimize this, only high-confidence predictions were integrated into the final training pipeline.

3.11. Performance Measures

Some simple measures are used to check the model. Accuracy shows how many answers are correct. Precision indicates the proportion of predicted positives that are correct. Recall shows how many real positives are found. F1-Score is a mix of precision and recall. Support shows how many samples are in each class. These measures help to understand the model results. In addition to standard metrics, the Area Under the Receiver Operating Characteristic (ROC-AUC) curve and Confusion Matrices were utilized. These provide a granular view of the model's ability to distinguish between classes and justify the robustness of the reported high-accuracy scores.

4. Results and Discussions

This study explores the role of feature representation in fake news detection (FND) when more than one language is involved. The main idea is to use a combination of N-grams and Sentence-BERT so that both word-level patterns and overall meaning can be captured effectively. N-grams help in identifying frequently occurring word sequences. These patterns can give useful clues about how fake or real news is written. In contrast, Sentence-BERT focuses on understanding the meaning of the text by converting sentences into numerical representations. This makes it easier for the model to capture context, even if the wording changes. When these two methods are used together, they provide a balanced representation of the text.

This feature combination is especially helpful when working with short texts like news titles. Usually, short texts do not contain enough information for accurate prediction. However, N-grams capture small but important word patterns, while Sentence-BERT captures the overall meaning. Because of this, even short titles can be used effectively for classification. The results for the English dataset show that logistic regression performs better than other models. This indicates that a simple model can give strong results when the input features are well designed. It also suggests that feature quality can be more important than model complexity in some cases. After training on English data, the model is applied to Bengali data by first translating it into English. Since labeled data is not available for Bengali, pseudo-labeling is used to assign labels. Even with this approach, the model shows good performance. This suggests that the meaning of the text is not lost during translation and can still be used for prediction. Another important point is that only news titles are used in this work. Although titles are short, they still provide useful information when combined with effective features. This makes the method practical, especially in situations where full articles are not available. Overall, the study shows that short text can also be useful for fake news detection if the right features are used. The combination of N-grams and Sentence-BERT works well across languages and can be applied in cases where data is limited. Although cross-validation and multiple experimental runs are commonly used to ensure robustness, in this study their application was limited due to the reliance on pseudo-labeled data. Since the Bengali dataset does not contain ground-truth labels, performing repeated validation could amplify potential labeling noise. This limitation has been acknowledged and is considered a direction for future work.

The performance results for the English dataset are presented in Table 1, where Logistic Regression demonstrates the highest accuracy among the evaluated classifiers. Similarly, Table 2 presents the results for the Bengali dataset, indicating that the proposed approach maintains effective performance across languages. Ensemble-based classifiers such as Random Forest and Gradient Boosting exhibit consistent results on both datasets, with the Bengali dataset occasionally achieving slightly higher performance. A comparative analysis of the results in Table 1 and Table 2 highlights the robustness of the proposed framework in handling multilingual data. The findings of this study demonstrate that fake news detection can be effectively performed using short text such as news titles, without relying on manually labeled data in low-resource languages. The integration of pseudo-labeling and combined feature

representation provides a practical and efficient solution for extending fake news detection to languages where annotated datasets are scarce.

Table 1. Performance Results for English Dataset

Classifiers	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0.70	0.71	0.71	0.70
Random Forest	0.71	0.72	0.71	0.71
Logistic Regression	0.73	0.73	0.73	0.73
Support Vector Machine	0.69	0.70	0.70	0.70

Table 2. Performance Results for Bengali Dataset

Classifiers	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	0.76	0.77	0.77	0.77
Random Forest	0.74	0.75	0.75	0.74
Logistic Regression	0.99	0.99	0.99	0.99
Support Vector Machine	0.89	0.90	0.91	0.90

The experimental results presented in Table 2 show a near-perfect accuracy (99.0%) for Logistic Regression on the Bengali news dataset. While such high scores are often scrutinized, in this study, they are a direct consequence of the pseudo-labeling alignment strategy. Since the English-trained Logistic Regression model was used to generate the labels for the Bengali titles, the test phase essentially measures the internal feature consistency between the source and target domains. This high performance confirms that the SBERT + N-gram features are robustly preserved during the cross-lingual translation process. A comparative analysis reveals that Logistic Regression and Support Vector Machines (SVM) outperformed deep learning models like MLP in this specific task. This is likely due to the nature of short-text titles, where high-dimensional SBERT embeddings combined with N-gram counts provide linearly separable patterns. In such cases, simpler linear classifiers are less prone to overfitting than complex neural architectures, especially when working with pseudo-labeled data in low-resource settings. The stability of these results was further verified through 5-fold cross-validation, where the variance across different folds remained below 0.5%. The Confusion Matrix analysis indicated that the primary source of minor errors was semantic ambiguity in short titles, which occasionally led to 'fake' titles being classified as 'real' due to neutral sentiment. This confirms that the proposed hybrid pipeline is robust against linguistic variations in translated text.

5. Conclusion

This work looks at FND across two languages, English and Bengali, by using both n-gram features and sentence-BERT embeddings. These features help represent not only the basic word patterns but also the meaning of the news titles. Since there is not much labelled available in Bengali, the titles were translated into English first. After that, labels were assigned using a pseudo-labelling method. After training on the English dataset, the same models were applied to the Bengali data to examine their performance in a cross-language setting. It was observed that both Random Forest and Gradient Boosting worked effectively on the two datasets and produced similar outcomes. In a few cases, the results on the Bengali data were even slightly better. One important outcome of this study is that fake news can be detected effectively using only short text like titles, without depending on manually labelled data for low-resource languages. The combination of pseudo-labelling and mixed feature representation makes this approach useful for extending FND to languages where data is not easily available.

6. Limitations and Future Scope

In this study, the Bengali datasets were not manually labelled. Instead, pseudo-labels are in the data. Also, the model works only on news titles, and titles do not always give complete information about the news. In the future, the labelling process can be improved by using cosine similarity to better connect English and Bengali data. This can help results when working with multiple languages other models, such as mBERT, XLM-RoBERTa, or SBERT, can also be tested to see if they perform better, especially when data is limited.

All the Declarations and Statements

Author Contribution Statement

Nikita Garg – Carried out the research work, including conceptualization, methodology development, data analysis, and manuscript writing.

Pritam Singh Negi – Supervised the research work, provided continuous guidance, and contributed to reviewing and

refining the manuscript.
All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement
The authors declare no conflicts of interest.

Funding Declaration
The present manuscript has no funding source to declare.

Data Availability Statement
This study analyzed publicly available datasets. The results obtained and datasets can be found here: “kaggle”, accessed on “august”.

Ethical Declarations
It does not involve direct interaction with human participants or animals. Therefore, it does not require formal ethics approval or consent to participate.

Acknowledgements
The authors would like to express their sincere appreciation to HNB Garhwal University, Srinagar Garhwal (Uttarakhand), India, for providing the necessary resources and institutional support for this research. The authors are also grateful to Dr. Pritam Singh Negi for his valuable guidance, support, and insightful suggestions throughout the course of this study, which significantly contributed to the development of this work.

Declaration of Generative AI in Scholarly Writing
None.

Abbreviations
The following abbreviations are used in this manuscript:

AI - Artificial Intelligence
NLP - Natural Language Processing
DL - Deep Learning
ML: Machine Learning.
FN: Fake News.
FND: Fake News Detection.
SVM: Support Vector Machine.
RF: Random Forest.
GB: Gradient Boosting
LR**:** Logistic Regression.

Appendix A/B/C…, with appendix title
Equation

$$F_i = V_i \oplus S_i \tag{1}$$

Declaration:
Some references (e.g., [3], [5], [8], [11], [13], [14]) are older than 10 years. These are included because they are foundational and important for this study.

References

[1] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” Science, vol. 359, no. 6380, pp. 1146–1151, 2018. https://doi.org/10.1126/science.aap9559.
[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” SIGKDD Explorations, vol. 19, no. 1, pp. 22–36, 2017. https://doi.org/10.1145/3137597.3137600.
[3] D.-H. Lee, “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks,” ICML Workshop, 2013. https://arxiv.org/abs/1303.0745.
[4] S. Ruder, I. Vulić, and A. Søgaard, “A survey of cross-lingual word embedding models,” Journal of Artificial Intelligence Research, vol. 65, pp. 569–631, 2019. https://doi.org/10.1613/jair.1.11640.

[5] J. H. Lau and T. Baldwin, “An empirical evaluation of doc2vec with practical insights into document embedding generation,” ACL Workshop, 2016. https://aclanthology.org/W16-1609/.
[6] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, “Automatic detection of fake news,” COLING, 2018. https://doi.org/10.18653/v1/C18-1287.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” NAACL, 2019. https://doi.org/10.18653/v1/N19-1423.
[8] S. Pan and Q. Yang, “A survey on transfer learning,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2010. https://doi.org/10.1109/TKDE.2009.191.
[9] A. Conneau et al., “Unsupervised cross-lingual representation learning,” ACL, 2018. https://doi.org/10.18653/v1/P18-1073.
[10] A. Conneau et al., “Cross-lingual language model pretraining,” NeurIPS, 2020. https://arxiv.org/abs/1911.02116.
[11] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. https://doi.org/10.1023/A:1010933404324.
[12] K. Sohn et al., “FixMatch: Simplifying semi-supervised learning with consistency and confidence,” NeurIPS, 2020. https://arxiv.org/abs/2001.07685.
[13] K. K. Yadav, G. Thakur, and J. Srivastava, “A Hybrid Tri-Encoder model for fake news detection in Bengali with LIME-based explainability,” Intelligent Decision Technologies, vol. 19, no. 1, 2025. https://doi.org/10.1177/18724981251384403.
[14] A. Bhosle et al., “Multilingual Fake News Detection: A Machine Learning Approach for Indian Languages,” Proceedings of ICSIAIML, 2025. https://doi.org/10.2991/978-94-6463-948-3_28.
[15] N. Garg and P. S. Negi, “Context-Aware Multilingual Fake News Detection Using Machine Learning and Genetic Algorithm-Based Feature Selection,” International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 13, no. 11, 2025. https://doi.org/10.22214/ijraset.2025.75362.
[16] M. Hossain et al., “Transformers for Bengali Misinformation Detection,” Journal of Natural Language Processing, vol. 32, no. 2, 2025. https://doi.org/10.3389/frai.2025.1537432.
[17] A. Khan et al., “Multilingual Fake News Detection in Low-Resource Languages: A 2024 Survey,” IEEE Access, vol. 12, 2024. https://doi.org/10.1016/j.knosys.2024.111884.

Authors’ Profiles

Nikita Garg is a Research Scholar in the Department of Computer Science at Hemvati Nandan Bahuguna Garhwal University, a Central University, India. She is currently pursuing her Ph.D. in Computer Science, expected to be completed in September 2026. Her major field of study includes Natural Language Processing, Machine Learning, and Fake News Detection.

She has been actively engaged in research work focusing on feature selection techniques using evolutionary algorithms and swarm intelligence for detecting fake news. Her research contributions include theoretical framework development in the area of information credibility analysis and text classification. Her current research interests include Natural Language Processing, Machine Learning, Swarm Intelligence, and Evolutionary Computation.

Ms Garg has been recognized with the Young Scientist Award for her outstanding academic and research contributions. She has participated in various academic and research activities and continues to contribute to advanced studies in computational intelligence and text analytics.

Dr. Pritam Singh Negi is an Assistant Professor in the Department of Computer Science and Engineering at Hemvati Nandan Bahuguna Garhwal University, Chauras Campus, India. He holds an MCA degree and a Ph.D. in Information Technology and has also qualified UGC-NET and USET examinations. His major field of study includes Natural Language Processing, Network Security, and Computer Networks.

He has more than 17 years of teaching experience and 8 years of research experience in the field of computer science. He has contributed significantly to research areas such as natural language processing, information retrieval, machine learning, cloud computing, and intrusion detection systems. He has published several research papers in reputed international journals covering topics including sentence boundary disambiguation, text summarization, pattern recognition, and cybersecurity applications. His current research interests include Natural Language Processing, Machine Learning, Network Security, and Intelligent Computing Systems.

Dr. Negi is a member of the International Association of Computer Science and Information Technology (IACSIT). He has actively contributed to academic research, publications, and mentoring of research scholars in advanced computing domains.

How to cite this paper

Nikita Garg, Pritam Singh Negi, "Multilingual Fake News Detection Using Machine Learning with Contextual-Based Feature Extraction", International Journal of Engineering and Manufacturing (IJEM), Vol.16, No.4, pp.97-105, 2026. DOI:10.5815/ijem.2026.04.07