International Journal of Applied Mathematics

Volume 38 No. 8s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

A COMPARATIVE STUDY OF MACHINE LEARNING AND DEEP LEARNING TECHNIQUES FOR MULTILINGUAL TEXT CLASSIFICATION

[1]Ami Shah, [2] Dr. Krunal kumar Narendrabhai Patel

[1]Computer Science & Engineering

Phd Scholar CVM University, Assistant Professor, Parul University,

Vadodara, India

amiphd22@gmail.com

[2]IT Department krunalkumar.patel@cvmu.edu.in

Associate Professor CVM University

Vadodara, India

Abstract

With the exponential increase in multilingual data generated through online platforms, the need for efficient and accurate text classification techniques has become more pressing than ever. Natural Language Processing (NLP) systems are now expected to handle linguistic diversity and cross-language semantics with greater precision.

This research aims to investigate the effectiveness of various machine learning and deep learning models in addressing the complexities of multilingual text classification. A carefully constructed multilingual dataset, encompassing diverse languages and scripts, was used as the basis for model training and evaluation.

The study involved a comparative analysis of several classification algorithms, including traditional approaches like Logistic Regression and Naïve Bayes, as well as advanced deep learning architectures. Specifically, models such as Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), and hybrid combinations like LSTM with Bidirectional GRU (BiGRU) and LSTM integrated with CNN layers were implemented.

Each model was trained and tested on the same dataset to ensure consistency in evaluation. The results revealed that hybrid deep learning models generally outperformed traditional machine learning classifiers. These models demonstrated superior precision, recall, and F1-scores, particularly in handling the nuances of multilingual content.

Furthermore, the findings offer valuable insights into how different models respond to challenges like class imbalance, a common issue in real-world datasets. The study not only highlights the strengths of deep learning techniques but also identifies areas for further optimization, such as improving recall for underrepresented classes.

Keywords : Machine Leaning, Deep Learning, CNN, LSTM

Received: August 04 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

1. Introduction

The widespread growth of digital communication across a variety of languages has created an urgent demand for reliable and accurate multilingual text classification systems. As online platforms continue to generate massive amounts of multilingual content, the limitations of conventional monolingual models have become increasingly apparent. These models, often trained on data from a single language, struggle to generalize across diverse linguistic structures and vocabularies.

To overcome this challenge, the development of multilingual processing models has emerged as a critical area of focus in Natural Language Processing (NLP). These models aim to recognize, interpret, and classify text data in multiple languages with minimal loss in accuracy.

This study presents a comprehensive evaluation of both traditional machine learning algorithms and cutting-edge deep learning architectures for multilingual text classification. Models such as Logistic Regression and Naïve Bayes are examined alongside more sophisticated architectures, including Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), and hybrid frameworks like LSTM+BiGRU and LSTM+CNN.

The experimental setup is centered on a labeled dataset comprising text entries categorized into two distinct classes. By comparing key performance metrics—accuracy, precision, recall, and F1-score—the study aims to reveal the strengths and weaknesses of each approach.

Multilingual text classification poses specific challenges, including language-specific semantic interpretation, syntactic variation across languages, and issues of data sparsity in less-resourced languages. This paper seeks to assess how effectively each model navigates these complexities, offering insights into their ability to generalize across linguistic boundaries.

2. Related Work

Recent literature has shown progress in multilingual NLP, leveraging transformer-based models and deep learning techniques. Traditional algorithms like Logistic Regression and Naïve Bayes are widely used for baseline comparisons due to their simplicity and efficiency on small datasets. In contrast, CNNs and LSTMs have demonstrated strength in learning sequential patterns and extracting semantic features.

Hybrid models combining different architectures (e.g., LSTM+BiGRU, LSTM+CNN) have also gained traction for their ability to enhance performance by capturing both local and long-term dependencies in textual data. However, their efficacy on multilingual datasets remains an area of active exploration.

3. Dataset Description

The dataset comprises text samples labeled into two binary classes: 0 and 1. It spans multiple languages, including English, Hindi, Bengali, and others. Preprocessing steps included:

- Tokenization

Received: August 04 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- Stop-word removal
- Text normalization (lowercasing, punctuation removal)
- Encoding and padding for deep learning models

The dataset contains **6,468** samples, with class distribution:

- Class 0: 4,947 samples
- Class 1: 1,521 samples

This imbalance necessitated careful metric evaluation, especially for recall and F1-score.

4. Methodology

4.1 Traditional Machine Learning Models

- **Logistic Regression:** A probabilistic binary classifier that uses a sigmoid function for output prediction.
- **Naïve Bayes:** A statistical classifier based on Bayes' Theorem with a strong independence assumption between features.

4.2 Deep Learning Models

- **CNN:** Captures local n-gram features using convolutional layers, ideal for semantic pattern recognition.
- **LSTM:** Specialized in sequence modeling with the ability to remember long-term dependencies.
- **LSTM+BiGRU:** A hybrid model that combines LSTM and bidirectional GRU layers to capture information from both past and future contexts.
- **LSTM+CNN:** Uses LSTM layers for temporal features and CNN layers for spatial/semantic feature extraction.

All deep learning models used embedding layers, dropout regularization, and the Adam optimizer. The models were trained on 80% of the data with 20% held out for testing.

5. Results and Evaluation

Received: August 04 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Performance metrics used:

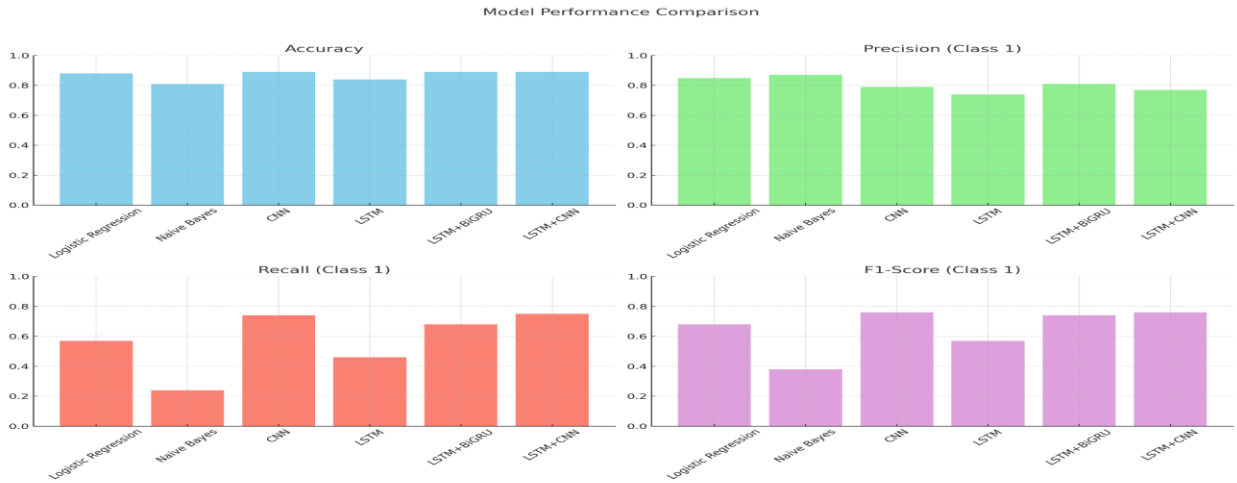
- **Accuracy**
  Precision
  Recall
  F1-score

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	**0.88**	**0.85**	**0.57**	**0.68**
Naive Bayes	**0.81**	**0.87**	**0.24**	**0.38**
CNN	**0.89**	**0.79**	**0.74**	**0.76**
LSTM	**0.84**	**0.74**	**0.46**	**0.57**
LSTM+BiGRU	**0.89**	**0.81**	**0.68**	**0.74**
LSTM+CNN	**0.89**	**0.77**	**0.75**	**0.76**

Table:1 Results of the dataset implementation on all the algorithms.

Observations:

- CNN and hybrid models (LSTM+BiGRU, LSTM+CNN) achieved the highest accuracy (0.89).
- Naïve Bayes struggled with class 1 recall due to its independence assumption and inability to handle contextual nuance.
- Logistic Regression provided a strong baseline but underperformed on minority class recall.



Received: August 04 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure:1 Graphical representation of all the results.

6. Discussion

The results suggest that deep learning models, particularly CNN-based and hybrid architectures, better understand multilingual semantics than traditional classifiers. The hybrid models benefit from both sequential and spatial features, enabling a more holistic understanding of text.

However, precision and recall for class 1 (minority) remained relatively lower than class 0 across models, indicating sensitivity to class imbalance. Addressing this issue through data augmentation, oversampling techniques, or cost-sensitive learning could further improve model fairness.

Additionally, while CNN-based models are faster to train and converge quickly, LSTM-based models provide better long-term context understanding, which is crucial for longer texts or code-mixed inputs.

7. Conclusion

This study demonstrates the effectiveness of deep learning, especially hybrid architectures, in handling multilingual text classification. While traditional models like Logistic Regression provide competitive accuracy, their limitations become apparent when dealing with class imbalance and contextual nuance.

The LSTM+BiGRU and LSTM+CNN models show promise for multilingual NLP tasks, offering a balanced trade-off between performance and computational cost. Future work may involve exploring transformer-based models like mBERT and XLM-R for more comprehensive multilingual understanding.

References

1. Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. Neural computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
2. Kim, Y. (2014). *Convolutional neural networks for sentence classification*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746–1751.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of NAACL-HLT 2019, 4171–4186.
4. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv preprint arXiv:1907.11692.
5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). *Unsupervised cross-lingual representation learning at scale*. Proceedings of ACL, 8440–8451.

Received: August 04 2025

Volume 38 No. 8s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

6. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.
7. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching word vectors with subword information*. Transactions of the Association for Computational Linguistics, 5, 135–146.
8. Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). *Bag of tricks for efficient text classification*. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 427–431.
9. Ruder, S., Vulić, I., & Søgaard, A. (2019). *A survey of cross-lingual word embedding models*. Journal of Artificial Intelligence Research, 65, 569–631.
10. Schwenk, H., & Douze, M. (2017). *Learning joint multilingual sentence representations with neural machine translation*. arXiv preprint arXiv:1704.04154.
11. Lison, P., & Tiedemann, J. (2016). *OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles*. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), 923–929.
12. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). *Hierarchical attention networks for document classification*. Proceedings of NAACL-HLT, 1480–1489.
13. Zhang, X., Zhao, J., & LeCun, Y. (2015). *Character-level convolutional networks for text classification*. Advances in Neural Information Processing Systems, 28, 649–657.
14. Kudo, T., & Richardson, J. (2018). *SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 66–71.
15. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation*. arXiv preprint arXiv:1609.0814

Received: August 04 2025