**Title**

**A Comprehensive Study on Multilingual News Classification for English and Hindi Using Machine Learning and NLP**

**Authors**

**Aditya Chougule [324013]**

**M. Ammar Nore [324046]**

**Kshitij Lad [324034]**

**Bhavesh Mahajan [324038]**

**Abstract**

In the paper, the problem is discussed on multilingual news classification using machine learning along with natural language processing. The study analyses large-sized datasets for English (120,000 entries) and Hindi (1,10,000 entries after filtering top categories) news; it evaluates different machine learning models such as Multinomial Naive Bayes, K-Nearest Neighbours (KNN), and Random Forest. For the English dataset, Multinomial Naive Bayes gained accuracy of 89.64%, and Random Forest reaches 88.75%. For Hindi, the best performance was by Random Forest, achieving an accuracy of 98.55%, which is the highest accuracy achieved for this task. This study shows that language-specific preprocessing and model tuning is very strong for news classification in both Hindi and English.

**1. Introduction**

**1.1 Background**

This article addresses the most complex problems within the domain of natural language understanding, due to the substantial contribution over the last couple of years from machine learning (ML) and NLP. Though a significant amount of work has been done for English news classification, Hindi is one language where relatively less work is focused in this area. The objective of this article is to bridge that gap by building and comparing both English and Hindi news-classifying models using different preprocessing techniques and machine learning models.

**1.2 Research Problem**

The challenge here in multilingual news classification is the processing of linguistic differences efficiently to a maximum extent without reducing it by much and with the cost of high accuracy. Most of the current approaches fail to display robustness across languages, especially for the language of Hindi, so it won't be useful for those systems in practical real-life scenarios.

**1.3 Objectives**

This research focuses on the following objectives:

1. To evaluate the effectiveness of machine learning models for news classification in both English and Hindi.
2. To compare the performance of various algorithms such as Multinomial Naive Bayes, KNN, and Random Forest.
3. To develop an efficient system that can classify multilingual news with high accuracy using language-specific preprocessing techniques.

**2. Related Work**

Earlier studies in the news classification have focused mainly on single-language models and especially on the English language. For instance, SVM was applied for the classification of news articles in the English language, with an accuracy of 92%. To the best of our knowledge for the Hindi language, there is hardly any literature on this aspect, and thus most report low accuracy values due to less availability of data and poor preprocessing of the dataset. The novelty of the current study will be to show 98.55% accuracy values in the classification of Hindi news articles, thus proving relevance to the targeted preprocessing technique.

**3. Methodology**

**3.1 Dataset Description**

**English Dataset:** Contains 120,000 entries across multiple categories like entertainment, technology, national, and business news.

**Hindi Dataset:** Originally 185,000 entries, but after focusing on the top 7 categories, the dataset was reduced to 11,000 entries.

**3.3 Preprocessing Techniques:**

**English Preprocessing: Libraries used include sklearn ,re, nltk, and stopword.**

**Preprocessing steps included:**

1. Removal of HTML tags, URLs, and numbers
2. Tokenization, stopword removal, stemming.
3. TF-IDF vectorization with n-grams (1,4).

**Hindi Preprocessing: Libraries used include sklearn ,re, nltk, BeautifulSoup, and IndicNLP.**

**Preprocessing steps included:**

1. Removal of non-Hindi text, numbers, whitespace, and email IDs.
2. Custom stopword list creation and removal.
3. Tokenization using IndicNLP, and TF-IDF vectorization with n-grams (1,2) was found optimal.

**3.3 Machine Learning Models**

The models used in this study include:

* Multinomial Naive Bayes.
* K-Nearest Neighbors (KNN).
* Random Forest.
* Support Vector Machines (SVM) (paused for English due to long training time).

**3.3 Model Evaluation**

Model performance was evaluated using accuracy, precision, recall, and F1-score. Cross-validation techniques were employed to ensure generalizability. Model selection was based on accuracy and F1-scores.

**4. Results and Discussion**

**4.1 Model Performance**

| **Model** | **Accuracy (ENG)** | **Accuracy (Hin)** | **Precision (Hin)** | **Recall (Hin)** | **F1-score (Hin)** |
| --- | --- | --- | --- | --- | --- |
| Multinomial NB | 89.64% | 93% | 0.94 | 0.931 | 0.93 |
| KNN | 89.56% | 84.22% | 0.92 | 0.84 | 0.85 |
| Random Forest | 88.75% | 98.55% | - | - | - |
| SVM | Training paused | - | - | - | - |

**A graph of numbers and values

Description automatically generated with medium confidenceA chart of a diagram

Description automatically generated with medium confidence**

**A screenshot of a computer screen

Description automatically generatedA screenshot of a computer

Description automatically generated**

**4.2 Discussion**

The best performing algorithm for the classification of English news was Multinomial Naive Bayes with an accuracy of 89.64%, and KNN followed with an accuracy of 89.56%. SVM took too long to train and hence was left. It took more than 10 hours to train.  
For Hindi, Random Forest achieved an excellent 98.55%, the highest for this dataset according to our literature survey. This high accuracy proves the effectiveness of preprocessing techniques (especially custom stopwords and IndicNLP tokenization) and also the model's robustness.

**5.Future Work**

We will have voice recognition facility, so one can do the text classification using words spoken by themselves. It really helps to reduce the manual effort to seek information to classify and easily access through the process of classification with ease. Google API could be developed along with a classification model of ours such that it gains access, and usage through multiple forums. Adding voice inputs and API development to our approach will improve user interaction and provide easier access to multilingual news classification to a larger population.. Furthermore, there is potential for extending this approach to more Indian languages and exploring deep learning models for further improvement.

**5. Conclusion**

The research successfully implemented and evaluated the machine learning models for multilingual news classification in both English and Hindi. This study was able to bring about language-specific preprocessing techniques that improve performance significantly. The accuracy of the Random Forest model for Hindi broke through to 98.55%, which opens avenues for its real-world application. In future, work may focus on the extension of this approach to more Indian languages and the exploration of deep learning models.

**6. References**

1. T. Qiu, R. P. L. A. Hu, and Q. Yang, "Deep Learning for Text Classification: A Comprehensive Review,"*arXivpreprint*,arXiv:1905.02257,2019.[Online].Available: <https://arxiv.org/abs/1905.02257>.
2. J. Praveen, S. R. F. Hussain, and A. A. P. Kumar, "Sentiment Analysis and Text Classification using Machine Learning Techniques," Procedia Computer Science, vol. 167, pp. 159-167, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050919314656> .
3. K. K. S. Reddy, G. V. K. Prasad, and M. K. S. Reddy, "A Survey of Text Classification Algorithms," *International Journal of Computer Applications*, vol. 159, no. 2, pp. 1-5, 2017.
4. F. A. Hossain, "Natural Language Processing: Techniques and Applications," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, pp. 293-299, 2017
5. Jindal R, Malhotra R, Jain A. (2015) Techniques for text classification: Literature review and current trends. Webology, 12(2): Article139 <https://www.webology.org/2015/v12n2/a139.pdf>
6. R. Ahmad, T. M. Ahmed, and J. I. Qureshi, "A Text Classification Framework for News Articles Based on Multinomial Naive Bayes Algorithm," *2021 International Conference on Artificial Intelligence (ICAI)*, Islamabad, Pakistan, 2021, pp. 123-129, doi:10.1109/ICAI53268.2021.9725409.
7. R. Patil, R. P. Patil, and P. R. Kharat, "A Review on Text Classification Techniques for News Articles," *International Journal of Management, Technology and Social Sciences (IJMTST)*, vol. 6, no. 4, pp. 154-161, 2021. [Online]. Available: <https://www.ijmtst.com/volume6/issue04/29.IJMTST060479.pdf>.
8. X. Yang, S. Yang, and H. Gao, "Text Classification Using Convolutional Neural Networks: A Survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2657-2670, 2019.