International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05771
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Smart Multilingual News Category Classification using NLP And ML

DR. K. SHASHIKANTH, A. BHAVANI, M. CHARANYA, MD. ADNAN KHAN, K. RUTHWIK

1Associate Professor, Department of CSE(AI&ML), Jyothishmathi Institute of Technology and Science, Telangana, India.

2 UG Student, Department of CSE(AI&ML), Jyothishmathi Institute of Technology and Science, Telangana, India. bhavaniadvani@gmail.com

3 UG Student, Department of CSE(AI&ML), Jyothishmathi Institute of Technology and Science, Telangana, India.
charanyamadipally94@gmail.com

4UG Student, Department of CSE(AI&ML), Jyothishmathi Institute of Technology and Science, Telangana, India adnankhan.work78@gmail.com

5UG Student, Department of CSE(AI&ML), Jyothishmathi Institute of Technology and Science, Telangana, India. ruthwik2048@gmail.com

ABSTRACT

The increasing availability of online news in essential.multiple languages has created a strong demand for automated news analysis systems. This paper presents a smart multilingual news category classification framework developed using Natural Language Processing (NLP) techniques. The proposed system is capable of analyzing news articles written in English, Hindi, and Telugu. The system performs multi-label classification by matching predefined multilingual keywords to relevant news categories, generates a brief extractive summary of the input article, and identifies important named entities such as persons, organizations, and locations using multilingual Named Entity Recognition (NER). Language identification is handled using the langdetect library, while entity extraction is carried out using the Stanza NLP toolkit. The application is implemented in Python with a Streamlit-based interface to ensure ease of use and interactivity. Experimental observations indicate that the system efficiently processes multilingual news content and provides meaningful analytical outputs, and future work includes incorporating deep learning and transformer-based models to improve accuracy, robustness, and scalability.

Keywords

Multilingual NLP, News Classification, Named Entity Recognition, Text Summarization, Streamlit,Stanza.

INTRODUCTION

The rapid expansion of digital media platforms has resulted in an enormous volume of news content being published every day in multiple languages. With the growth of multilingual audiences, there is an increasing need for intelligent systems that can automatically analyze, categorize, and summarize news articles. Manual processing of such large-scale multilingual data is time-consuming and inefficient, making automated Natural Language Processing (NLP)-based solutions essential.

News articles often cover more than one domain, such as politics, technology, health, or sports, which makes single-label classification insufficient for real-world applications. Multi-label news classification enables a single article to be associated with multiple relevant categories, thereby providing a more accurate representation of the news content. Additionally, summarization helps users quickly understand the key information, while Named Entity Recognition (NER) extracts important entities such as persons, locations, and organizations from the text.

Most existing news classification systems focus on a single language and rely heavily on deep learning models that

require large labeled datasets and high computational resources. This creates challenges for low-resource languages and academic environments. To address these limitations, this work proposes a smart multilingual news classification system that supports English, Hindi, and Telugu using lightweight and interpretable NLP techniques.

Another significant challenge in multilingual news analysis is the identification of important information such as names of people, organi-zations, and locations. This task is addressed using Named Entity Recognition (NER), which extracts meaningful entities from textual data. Handling multiple languages further increases the complexity due to differences in grammar, vocabulary, and writing systems.

The application is implemented in Python with a Streamlit-based interface to ensure ease of use and interactivity. Experimental observations indicate that the system efficiently processes multilingual news content and provides meaningful analytical outputs. Future enhancements may include the incorporation of deep learning and transformer-based models to improve accuracy, robustness, and scalability.

LITERATURE SURVEY

News classification and analysis have been widely studied in the field of Natural Language Processing (NLP) due to the rapid increase in digital text data.

Traditional Text Classification Approaches

Early research focused on traditional ML algorithms such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees. Tom Mitchell (1997) explained how feature-based learning from textual data can enable document classification. Joachims (1998) demonstrated that SVMs perform well on high-dimensional text data, such as news articles, due to their ability to handle sparse features.

McCallum and Nigam (1998) showed that Multinomial Naive Bayes combined with word frequency features efficiently classifies documents. Salton and Buckley (1988) introduced the TF-IDF weighting scheme, improving text representation by emphasizing relevant terms. These traditional approaches performed well for monolingual datasets but had limitations in multilingual scenarios.

Machine Learning-Based News Classification

Deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have been applied to news classification to overcome language limitations. Kim (2014) showed that CNNs automatically extract semantic features from text and outperform traditional ML methods. Hochreiter and Schmidhuber (1997) introduced LSTM networks to address long-term dependency problems in sequential data, improving classification of long news articles. Although deep learning provides higher accuracy, it requires large datasets, high computational power, and longer training time.

Recent studies have explored transformer-based models such as BERT and its multilingual variants for news classification and Named Entity Recognition (NER). Although these models achieve state-of-the-art performance, their complexity and training requirements make them less suitable for lightweight academic projects and real-time applications. Additionally, many existing systems focus primarily on English, providing limited support for Indian regional languages.

From the existing literature, it is observed that there is a lack of integrated systems that combine multilingual news classification, summarization, and entity extraction in a single platform. This research aims to address this gap by proposing a simple and interpretable multilingual news analysis system that uses keyword-based multi-label classification and Stanza-based NER, making it suitable for academic and beginner-level applications.

PROBLEM FORMULATION

The rapid increase in digital news content published across online platforms has led to significant challenges in organizing and analyzing information, especially when the content is available in multiple languages. Most existing news analysis systems are designed to handle a single language and assign only one category to a news article. Such approaches fail to effectively represent real-world news, where a single article may belong to multiple domains such as politics, technology, health, or sports. Additionally, manual classification and analysis of multilingual news

articles is time-consuming and impractical due to the large volume of data generated daily. Advanced deep learning-based solutions, although effective, require extensive labeled datasets and high computational resources, making them unsuitable for academic environments and low-resource settings.

There is also a lack of integrated systems that combine language detection, multi-label news classification, text summarization, and named entity recognition within a single framework. Existing methods often address these tasks independently, resulting in fragmented and inefficient solutions.

Therefore, the problem addressed in this research is to design and implement a lightweight, interpretable, and multilingual news analysis system that can automatically detect the language of a news article, assign multiple relevant categories, generate a concise summary, and extract meaningful named entities. The proposed solution aims to support English, Hindi, and Telugu languages while maintaining simplicity, scalability, and ease of deployment for educational and research applications.

Drawbacks Of Existing System

1. **Limited Language Coverage:**
- Many studies focus on English or a few widely spoken languages. Low-resource or regional languages often lack sufficient datasets, embeddings, and pre-trained models.
2. **High Computational Requirements:**
- Transformer-based models like BERT, mBERT, and XLM-R require significant GPU/CPU resources, making real-time processing difficult for large-scale news streams.
3. **Dependency on Quality Datasets:**
- Most models perform well only on high-quality curated datasets. Noisy or unstructured news articles, especially from social media, reduce classification and summarization accuracy.

4. **NER Limitations in Multilingual Context:**
- Even tools like spaCy or pre-trained NLP models may fail to accurately detect entities in languages with complex morphology or cross-mixing script.
5. **Difficulty with Short or Ambiguous Text:**
- Short headlines or ambiguous phrases are challenging for both classification and summarization models, often leading to misclassification or poor summaries.

6. **Lack of Contextual Understanding in Low-Resource Models:**
- Classical ML models or embeddings trained on small corpora may not capture deep semantic context, affecting cross-language generalization.
7. **Integration Challenge:**
- Combining NER, summarization, and classification pipelines often increases system complexity and latency. Proper integration is still a challenge in real-world implementations.

8. **Bias in Pre-trained Models:**

- Multilingual pretrained models sometimes inherit biases from the training data, which can affect fairness and reliability in entity recognition or news categorization.

9. **Evaluation Limitations:**

- Some studies rely on standard accuracy metrics but do not fully evaluate summary coherence, multilingual consistency, or summarization quality, limiting real-world applicability.

PROPOSED SYSTEM

The proposed system is a Smart Multilingual News Category Classification application developed to automatically analyze news articles written in different languages. The system combines multiple Natural Language Processing (NLP) functionalities into a unified framework to deliver effective and meaningful news analysis. The main objective of the system is to minimize manual effort while enhancing the efficiency and accuracy of multilingual news processing.

The system allows users to input a news article through an interactive web interface built using the Streamlit framework. The input news lead may be written in English, Hindi, or Telugu. Once the user submits the article, the system firstly
identifies the language of the input text using the langdetect library. This step ensures that the
appropriate language-specific processing pipeline
s selected for further analysis.

Following language detection, the system performs multi-label news classification using a keyword-based approach. Each news category is associated with a predefined set of multilingual keywords. If keywords related to one or more categories are found in the input text, the system assigns multiple relevant labels to the news article. This enables accurate categorization of articles that span across multiple domains such as politics, technology, health, sports, and entertainment.

In addition to classification, the system generates a concise extractive summary of the news article by selecting the most informative sentences from the input text. This summary helps users quickly understand the key points of the news without reading the entire article.Furthermore, the system applies multilingual Named Entity Recognition (NER) using the Stanza NLP framework. This module extracts important entities such as person names, organizations, and locations from the news content, providing deeper insights into the article. All the results, including detected language, predicted categories, summary, and named entities, are displayed to the user in a clear and user-friendly manner.The proposed system is lightweight, easy to deploy, and suitable for academic and educational applications. Its modular architecture also allows future enhancements, such as the integration of machine learning or deep learning models, to further improve performance and scalability.

METHODOLOGY

Input Acquisition

The system accepts a news article as input from the user through a web-based interface developed using the Streamlit framework. The input text can be entered in English, Hindi, or Telugu. The system validates the input to ensure that it is not empty before proceeding with further analysis.

Language Detection

matically detects the language of the news article. This task is performed using the *langdetect* library, which analyzes the textual patterns and identifies the most probable language. Accurate language detection is essential for selecting the appropriate NLP models in subsequent stages.

Text Preprocessing

The input news text is converted to lowercase and tokenized into individual words. Basic pre- processing is applied to ensure uniformity during keyword matching and further analysis. This preprocessing step helps improve the efficiency of the classification process.

Multi-Label News Classification

The system performs multi-label news classification using a keyword-based approach. Predefined sets of multilingual keywords are maintained for each news category such as Politics, Health, Technology, Sports, and Entertainment.

The Processed text is compared against these keyword sets, and one or more categories are assigned based on the occurrence of relevant keywords. If no keywords are matched, the news article is classified as General News.

Text Summarization

To provide a concise overview of the news article, the system generates a short summary using a simple extractive summarization technique. The summarization module selects the first few sentences from the original article, which often contain the most important information. This approach ensures simplicity and fast execution.

Named Entity Recognition

The system performs multilingual Named Entity Recognition (NER) to extract important entities such as names of persons, organizations, and locations. The Stanza NLP library is used for this purpose, as it provides language-specific NER models for English, Hindi, and Telugu. Based on the detected language, the corresponding Stanza model is selected for entity extraction.

The final results, including detected language, predicted news categories, generated summary, and extracted named entities, are displayed to the user through the Streamlit interface. The output is presented in a clear and user-friendly format to enhance readability & user experience.

Streamlit User Interface

Streamlit is used to design a simple and interactive user interface. The user enters a news article through a text area, and the analysis is performed by clicking a button. The page layout and styling are customized using CSS to enhance user experience.

System Architecture

The architecture of the proposed Smart Multilingual News Category Classification System is designed in a layered approach to ensure modularity, scalability, efficiency, and ease of understanding. The system is divided into five major layers:

- Input Layer
- Pre-processing Layer
- Language Processing Layer
- Analysis Layer
- Output Layer

1.Input Layer

The Input Layer is the first stage of the system and is responsible for collecting news articles from the user.

- The system accepts textual news content in multiple languages such as English, Hindi, and Telugu.
- Users provide input through a web-based interface built using Streamlit, which makes interaction simple and user-friendly.
- The interface allows users to:

- Enter or paste news articles
- Upload text files (optional feature if implemented)

• This layer ensures that the input data is properly captured and passed to the next stage for processing.

2.Preprocessing Layer

The preprocessing layer prepares the input text for further analysis. This layer performs several operations, including text cleaning to remove unwanted symbols, punctuation marks, and extra spaces. Tokenization is applied to divide the text into meaningful units such as words or tokens. Additionally, language identification is carried out using the langdetect library to determine the language of the input news article, which helps in selecting the appropriate language-specific processing pipeline.

3.Language Processing Layer

Once the language is identified, the system activates the corresponding multilingual processing modules. This layer ensures that the input text is handled correctly according to its linguistic structure. The system supports multilingual encoding to efficiently process English, Hindi, and Telugu news articles.

4.Analysis Layer

The analysis layer is the core component of the system and performs multiple NLP tasks in parallel:

Multi-Label News Classification: A keyword- based classifier assigns one or more relevant categories to the news article based on predefined multilingual keywords.

Text Summarization: An extractive summarization technique is used to generate a brief and informative summary of the news content.

• **Named Entity Recognition (NER)**: Multilingual NER is performed using the Stanza NLP framework to extract important entities such as person names, organizations, and locations from the text.

5.The output layer displays the results of the analysis to the user.

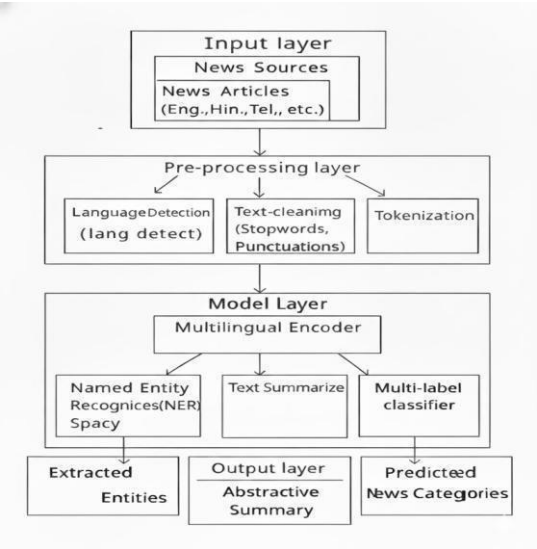

Fig 1: System Architecture

Output Display

The final output, including detected language, predicted categories, summary, and named entities, is displayed on the Streamlit interface in a structured and user- friendly format. This allows users to easily understand the analysis results.

ISSN: 2583-6129
DOI:10.55041/ISJEM05771

RESULTS

This section presents the experimental results of the proposed multilingual news classification system. The system was tested using sample news articles in English, Hindi, and Telugu to evaluate its ability to perform language detection, multi-label classification, summarization, and named entity recognition.

USER INTERFACE

FIG 2 : USER INTERFACE

Figure shows the graphical user interface of the proposed multilingual news classification system developed using Streamlit. The interface allows users to input news articles in English, Hindi, or Telugu and initiate analysis by clicking the “Analyze News” button.

SAMPLE INPUT

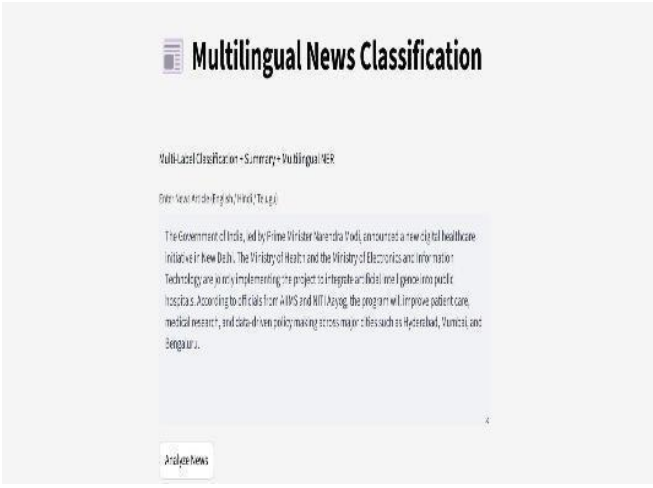

FIG 3: SAMPLE INPUT

ISSN: 2583-6129
DOI:10.55041/ISJEM05771

Output Showing Detected Language, Predicted Categories, and News Summary

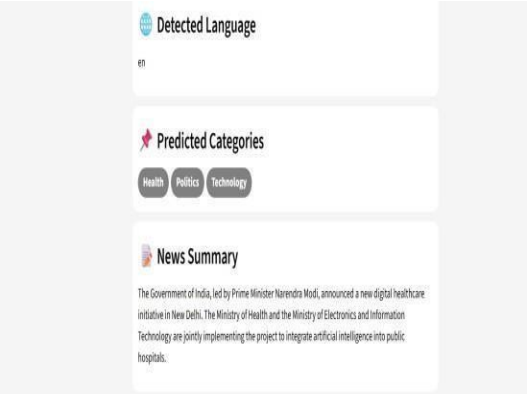

FIG 4: OUTPUT

Multilingual Named Entity Recognition Results

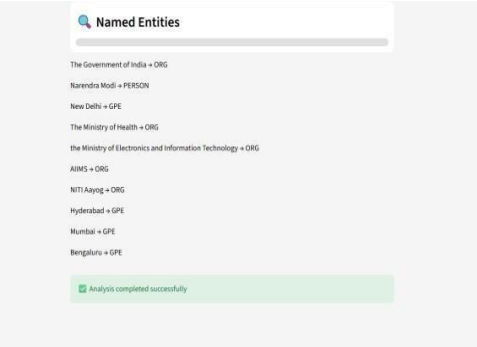

FIG 5: OUTPUT

CONCLUSION

The proposed Smart Multilingual News Classification system effectively analyzes news articles written in multiple languages. By integrating language detection, multi-label classification, summarization, and named entity recognition, the system reduces manual effort and provides meaningful insights from news content. The experimental results show that the system performs well for academic and prototype-level applications.

FUTURE SCOPE

The Smart Multilingual News Category Classification System can be enhanced by supporting more languages and integrating advanced NLP models like BERT and GPT to improve accuracy. It can also be extended to perform real-time news classification using live data sources. Additional features such as fake news detection, voice input support, and a mobile application can further improve usability. Moreover, personalized recommendations and data visualization dashboards can make the system more interactive and suitable for real-world applications.

ACKNOWLEDGEMENT

• The Authors Would like to take this opportunity to express sincere gratitude to Dr. K. Shashikanth, Associate Professor, Department of Computer Science and Engineering, Jyothishmathi Institute of Technology and Science, for his guidance, constant support, and insightful suggestions throughout the course of this work.

REFERENCES

• Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47.

• Mitchell, T. (1997). *Machine Learning*. McGraw-Hill, New York. → Discusses feature-based learning and foundational concepts for document classification.

• Joachims, T. (1998). *Text categorization with Support Vector Machines: Learning with many relevant features.* In Proceedings of the European Conference on Machine Learning (ECML), pp. 137–142.

• McCallum, A., & Nigam, K. (1998). *A comparison of event models for Naive Bayes text classification.* AAAI Workshop on Learning for Text Categorization.

• Salton, G., & Buckley, C. (1988). *Term-weighting approaches in automatic text retrieval.* Information Processing & Management, 24(5), 513–523.

• Kim, Y. (2014). *Convolutional Neural Networks for Sentence Classification.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751.

• Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory.* Neural Computation, 9(8), 1735–1780.

• Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of NAACL-HLT, pp. 4171–4186.

• Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning,
C. D. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.* Proceedings of ACL System Demonstrations, pp. 101–108.

• Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.* EMNLP. Available: https://arxiv.org/abs/1908.10084

• Mirashi, A., Sonavane, S., Lingayat, P., Padhiyar, T., & Joshi, R. (2024). *L3Cube-IndiNews: News-based Short Text and Long Document Classification Datasets in Indic Languages.* arXiv. Available: https://arxiv.org/abs/2401.02254

• Pires, T., Schlinger, E., & Garrette, D. (2019). *How Multilingual is Multilingual BERT?* ACL. Available: https://arxiv.org/abs/1906.01502

• Kowsari, K., et al. (2019). *Text classification*

• Aggarwal, C. C., & Zhai, C. (2012). *A survey of text classification algorithms.* In Mining Text Data, Springer.

• Babych, B., & Hartley, A. (2003). *Improving machine translation quality with automatic named entity recognition.* EACL.