

Automated Crawling, Categorization and Sentiment Analysis of Digital News with Incorporated Feedback System

K Aishwarya¹, Kuchi Susmitha², Megha H U³, Meghana K⁴, Prof. Narendra Mani Tripathi⁵
^{1,2,3,4,5}Department of Computer science and engineering, Reva University, Bangalore, India

Abstract— The rapid growth of digital news platforms has increased the difficulty of monitoring large volumes of information published across multiple sources and languages. Manual tracking of such information is time-consuming and often inefficient, particularly when timely identification of important developments is required. This paper presents an automated multilingual news monitoring system designed to collect, categorize, and analyze digital news articles using natural language processing techniques and machine learning methods.

The proposed system retrieves news articles automatically from multiple RSS feed sources and processes them through a structured analysis pipeline. Sentence-transformer embeddings combined with logistic regression are used to categorize articles into thematic domains, while sentiment polarity is evaluated using the VADER sentiment analysis model. The system further identifies severity levels based on sentiment scores and groups related articles describing the same events using agglomerative clustering. A notification mechanism generates automated alerts when high-severity negative news is detected. An interactive dashboard is also provided to visualize sentiment trends, category distribution, and geographic information extracted from news articles.

The experimental results demonstrate that the proposed system provides an efficient framework for multilingual news monitoring and supports faster identification of critical developments through automated analysis and visualization.

Keywords: News Crawling, Natural Language Processing, Sentiment Analysis, Multilingual Classification, Event Clustering, News Monitoring System

I. INTRODUCTION

Digital news media plays an important role in shaping public awareness and supporting decision-making processes across government organizations, research institutions, and industries. With the increasing number of online news sources publishing content continuously, it has become difficult to manually monitor large

volumes of information efficiently. This challenge becomes more complex in multilingual environments where relevant information may appear in different regional languages.

Traditional approaches for monitoring digital news rely mainly on manual browsing or simple keyword-based filtering methods. These approaches are often insufficient for identifying emerging developments in real time and may overlook important regional information. Automated systems that combine natural language processing and machine learning techniques provide an effective solution for addressing these limitations.

The system proposed in this work automatically collects news articles from multiple RSS feed sources and processes them using a structured pipeline consisting of preprocessing, categorization, sentiment evaluation, clustering, and visualization modules. Articles with high-severity negative sentiment are identified and reported through automated alert notifications. The system also provides an interactive dashboard that enables users to explore trends and filter results based on language, category, and sentiment.

This approach improves the efficiency of monitoring multilingual news streams and supports timely identification of critical developments.

II. RELATED WORK

Automated news monitoring systems have been widely studied due to the increasing availability of digital information across online platforms. Earlier research in this area focused mainly on web scraping techniques for collecting news articles from static and dynamic

web pages. Tools such as RSS feed parsers and structured content extraction libraries have been used to improve the reliability of news collection.

Machine learning techniques including Support Vector Machines, Naïve Bayes classifiers, and logistic regression models have been applied for news classification tasks. More recently, transformer-based embedding models have been introduced to improve semantic understanding of multilingual text content.

Sentiment analysis techniques have also been widely used to determine the tone of news articles. Lexicon-based sentiment analyzers such as VADER provide efficient polarity estimation without requiring extensive training datasets. These techniques are particularly useful for identifying news articles that may represent critical developments.

Although several systems have been proposed for automated news monitoring, many existing solutions focus only on single-language processing and do not provide integrated clustering and alert generation mechanisms. The proposed system addresses these limitations by supporting multilingual analysis, duplicate removal, event clustering, and automated feedback notification within a unified framework.

III. PROPOSED MODEL

The proposed model for the Automated Crawling, Categorization, and Sentiment Analysis of Digital News with Incorporated Feedback System is a robust, modular framework that automates the entire pipeline—from collecting news data to delivering actionable insights to government departments. This end-to-end system combines advanced web scraping, multilingual processing, machine learning classification, sentiment evaluation, and real-time alerting, ensuring efficient and scalable analysis of digital news.

1. News Crawling Module

The initial stage involves automated news data collection from a wide variety of online news portals. The system employs:

- feedparser for retrieving news articles from multiple RSS feed sources and extracting structured metadata such as headlines, publication dates, summaries, and source links.
- trafilatura for extracting the main textual content of

news articles from web pages while removing advertisements, navigation elements, and unrelated page components.

- hashlib (SHA-256 hashing) for detecting and eliminating duplicate articles collected from different sources, ensuring that repeated news entries are not stored multiple times in the database.

This crawling mechanism supports news in multiple languages, including English, Hindi, Tamil, Telugu, and Kannada, thereby improving the accessibility and effectiveness of the system for multilingual news monitoring.

2. Text Preprocessing Module

Raw news articles collected from online sources often contain unwanted formatting elements such as HTML tags, scripts, and special characters. Therefore, preprocessing is performed to clean and normalize the textual content before further analysis. The system employs:

- BeautifulSoup for removing HTML tags and extracting clean textual content from article pages.
- spaCy natural language processing library for performing tokenization, stopword removal, and lemmatization on English-language news articles to improve text consistency and readability.
- spaCy Named Entity Recognition (NER) for identifying important entities such as persons, organizations, and geographic locations mentioned in the news articles, which supports downstream visualization and geo-mapping features.

For regional language articles, basic normalization techniques such as whitespace cleaning and text filtering are applied to prepare the content for multilingual embedding-based classification.

3. Categorization Module

The categorization module assigns each news article to an appropriate thematic domain such as politics, health, crime, education, technology, or environment. This classification helps organize news content into meaningful groups for easier analysis. The system

employs:

- sentence-transformers library using the *paraphrase-multilingual-MiniLM-L12-v2* model to convert multilingual news text into semantic embedding vectors that capture contextual meaning across languages.
- Logistic Regression classifier (scikit-learn) for predicting the most relevant category based on generated embedding features.
- Cosine similarity (scikit-learn) as a fallback mechanism for assigning categories when classifier confidence is below the predefined threshold, ensuring reliable classification even for unseen content.

This hybrid classification strategy improves accuracy and robustness when processing multilingual news articles.

4. Sentiment Analysis Module

Sentiment analysis is performed to determine the overall tone of each news article and identify whether the content expresses positive, neutral, or negative sentiment. This helps in detecting potentially sensitive developments reported in digital news sources. The system employs:

- VADER (Valence Aware Dictionary and Sentiment Reasoner) through the vaderSentiment library for computing sentiment polarity scores without requiring additional training data.
- Compound sentiment score evaluation to classify articles into positive, neutral, or negative sentiment categories based on predefined threshold values.

This rule-based sentiment evaluation approach provides efficient and reliable polarity detection for news-style textual content.

5. Severity Detection Module

The severity detection module evaluates the importance of news articles based on their sentiment polarity scores. Articles with stronger negative sentiment values are considered more critical and require higher attention. The system employs:

- Sentiment score threshold mapping derived from VADER compound scores to classify articles into severity levels such as Critical, High, Medium, and None.
- Priority-based filtering mechanism to highlight high-severity articles for monitoring and alert generation within the dashboard interface.

This severity classification improves the ability of the system to identify potentially important developments quickly.

6. Event Clustering Module

Multiple news sources often report the same event with slight variations in wording. To reduce redundancy and improve interpretability, the system groups related articles describing the same event into clusters. The system employs:

- TF-IDF vectorization (scikit-learn) for converting news article text into numerical feature representations based on term importance.
- Agglomerative clustering algorithm (scikit-learn) for grouping semantically similar articles into event-based clusters without requiring predefined cluster counts.

This clustering mechanism allows users to analyze multiple perspectives of the same event more efficiently.

7. Feedback and Notification Module

To support timely identification of important developments, the system includes an automated notification mechanism that alerts users when high-severity negative news articles are detected. The system employs:

- SMTP-based email notification service for sending automated alerts containing article summaries, sentiment classification, and severity information.

- Severity-level filtering logic to ensure that alerts are generated only for articles classified under high and critical risk categories.

This real-time feedback mechanism improves the usefulness of the system by enabling faster response to sensitive news events without requiring continuous manual monitoring.

Modular and Scalable Architecture

The proposed system follows a modular architecture in which each component performs a specific role within the overall processing pipeline. This structure allows individual modules to be improved or extended without affecting the functionality of the remaining system. As a result, the system can be adapted easily to support additional analytical capabilities in future versions. For example:

- The sentiment analysis module currently uses the VADER sentiment analysis model, which can later be enhanced by integrating transformer-based multilingual models such as BERT-based sentiment classifiers for improved performance on regional language news content.
- The multilingual processing capability of the system can be extended further by incorporating additional embedding models supported by the sentence-transformers framework, enabling analysis of a wider range of regional and international news sources.
- The existing RSS-based news collection pipeline can be expanded to include additional real-time information streams such as domain-specific monitoring feeds or verified social media sources, improving coverage and responsiveness of the system.

This modular structure improves flexibility and ensures that the system remains adaptable for future enhancements while maintaining a clear workflow from automated data collection to structured analysis and visualization.

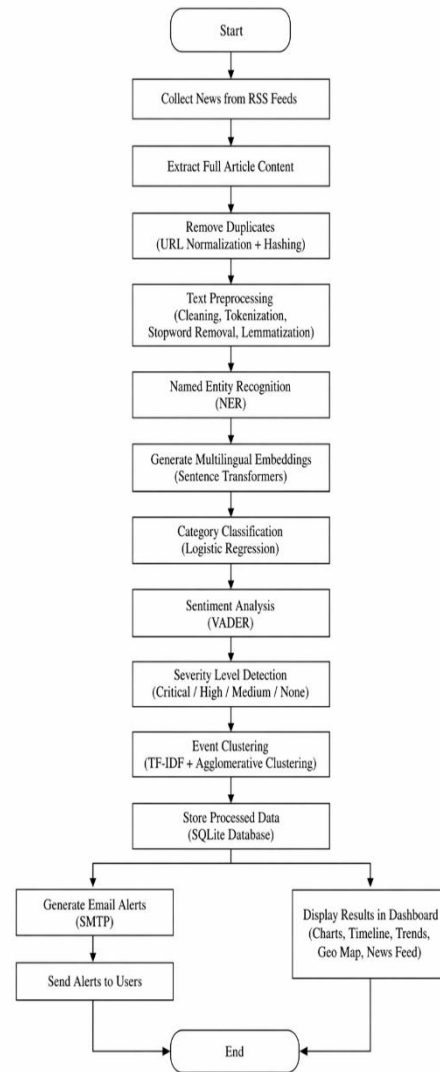


Fig. 1. Flowchart

IV. RESULT

The proposed system successfully collected multilingual news articles from multiple RSS feed sources covering both national and regional media platforms. The crawling module efficiently extracted article metadata and full textual content for further processing.

- **Accurate Categorization:**

News articles were categorized into thematic domains such as politics, health, environment, crime, and technology using sentence-transformer embeddings with logistic regression classification.

- **Sentiment Analysis Performance:**

Sentiment polarity was determined using the VADER

sentiment analysis model, which helped identify positive, neutral, and negative articles and supported severity-level detection.

- **Duplicate Detection Efficiency:**
A three-layer duplicate detection approach based on URL normalization and hashing techniques reduced repeated entries and improved dataset quality.
- **Real-Time Notifications:**
High-severity negative news articles automatically triggered email alerts, enabling timely monitoring of important developments.
- **Interactive Dashboard Visualization:**
The dashboard provided filtering options and visual representations such as charts and geo-mapping to support easier interpretation of news trends.
- **Event Clustering Capability:**
Agglomerative clustering with TF-IDF representation grouped related articles describing the same event, reducing redundancy.
- **Multilingual Support:**
The system processed news content across multiple languages including English, Hindi, Tamil, Telugu, and Kannada, improving coverage of regional information sources.



Fig .2. Dashboard of Govintel website

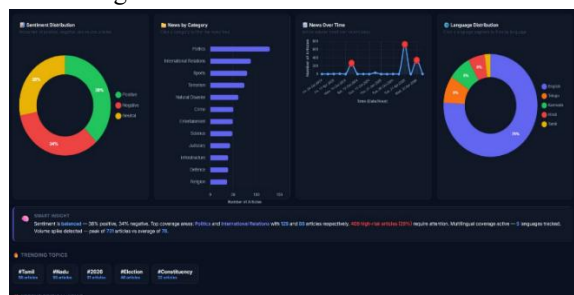


Fig .3. News analytics dashboard overview.

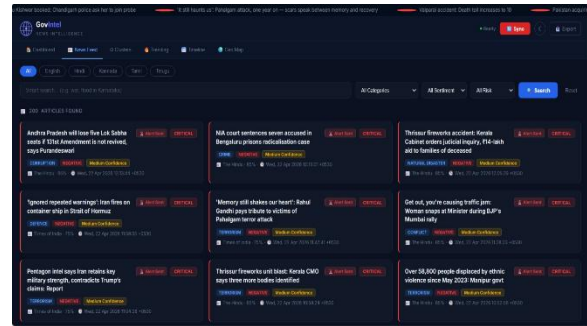


Fig.4. News monitoring interface with risk labels.

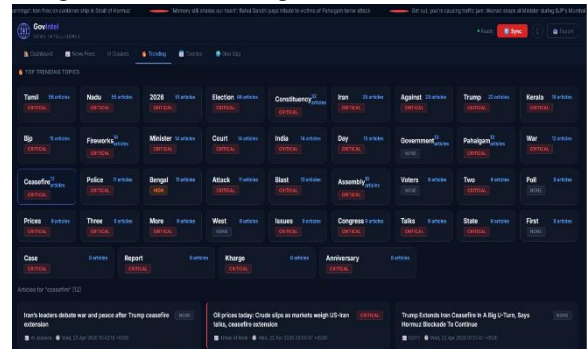


Fig.5.Display of Trending news

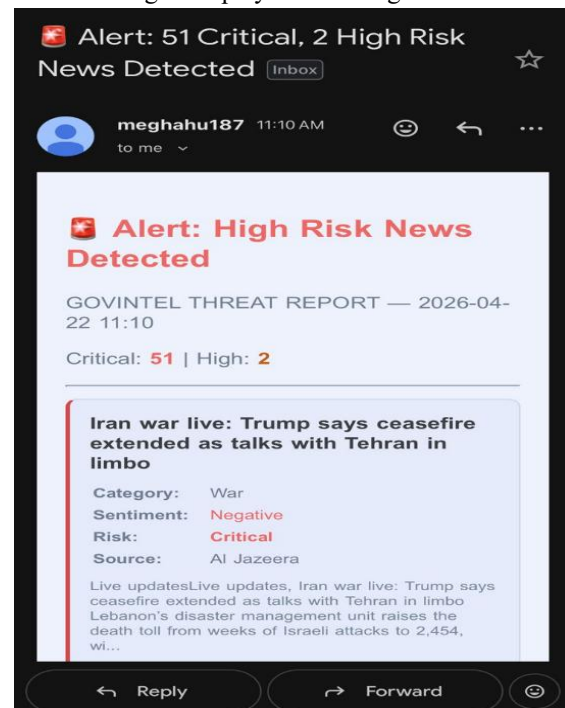


Fig .6 . Receiving alert mail

V. CONCLUSION AND FUTURE SCOPE

This project presented an automated multilingual news monitoring system capable of collecting, categorizing, and analyzing digital news articles using natural

language processing and machine learning techniques. The system integrates RSS-based crawling, preprocessing, semantic categorization, sentiment evaluation, clustering, and automated alert generation within a unified processing framework.

By combining multilingual embedding models with sentiment-based severity detection and clustering techniques, the system improves efficiency in monitoring large volumes of news content across multiple sources and languages. The interactive visualization dashboard further enhances usability by allowing users to explore trends and filter results based on multiple analytical parameters.

Future improvements may include integration of advanced multilingual transformer-based sentiment models to improve classification accuracy across regional language news content. Additional extensions such as mobile application support, enhanced geo-mapping coverage, and integration with verified real-time information sources can further strengthen the system's capability for large-scale news intelligence monitoring.

REFERENCE

- [1] Smart India Hackathon 2023, Problem Statement 1329, "360-degree feedback software to monitor Government of India-related news stories in regional media using Artificial Intelligence and Machine Learning," Available: <https://www.sih.gov.in>, Accessed: 2023.
- [2] P. K. Singh, S. Kumar, and R. Verma, "Multilingual News Classification Using Deep Learning Models," *International Journal of Computational Linguistics and Natural Language Processing*, vol. 10, no. 2, pp. 45–58, 2024.
- [3] J. Zhao, M. Liu, and L. Chen, "Real-time Sentiment Analysis of Social Media Data for Public Opinion Monitoring," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 1, pp. 123–134, Jan. 2024.
- [4] R. Verma and N. Singh, "An Efficient Framework for Multilingual News Crawling and Sentiment Analysis," *Proceedings of the International Conference on Artificial Intelligence and Data Engineering (AIDE)*, 2023, pp. 150–160.
- [5] M. A. Rahman, S. K. Debnath, and T. Ahmed, "Deep Learning Approaches for Multilingual Text Classification: A Review," *IEEE Access*, vol. 11, pp. 34350–34368, 2023.
- [6] H. Kim and J. Park, "A Real-Time News Aggregation and Sentiment Analysis System for Financial Markets," *IEEE Transactions on Big Data*, vol. 9, no. 4, pp. 987–996, 2023.
- [7] A. T. Nguyen, Q. H. Tran, and P. M. Le, "Multilingual News Text Classification Using Transfer Learning," *International Journal of Data Science and Analytics*, vol. 16, pp. 301–312, 2024.
- [8] S. Lee and M. Jeong, "Automatic Categorization of News Articles Using Machine Learning and NLP," *Proceedings of the ACM Symposium on Applied Computing*, 2023, pp. 712–719.
- [9] S. K. Patel and A. K. Jha, "An Overview of Web Scraping Tools and Techniques for News Aggregation," *International Journal of Information Technology*, vol. 15, no. 3, pp. 677–684, 2023.
- [10] V. Singh and M. Singh, "Multilingual Sentiment Analysis: Challenges and Approaches," *Journal of Artificial Intelligence Research*, vol. 78, pp. 921–945, 2023.
- [11] J. Park and H. Lee, "Real-Time News Sentiment Analysis System Using Deep Learning," *Applied Soft Computing*, vol. 130, 2023, Art. no. 109646.
- [12] M. Kumar and R. S. Yadav, "Machine Learning Techniques for News Classification: A Comparative Study," *Journal of Big Data Analytics*, vol. 7, no. 2, pp. 115–126, 2023.
- [13] N. Sharma and V. Gupta, "Design and Implementation of a Real-Time News Monitoring System," *Proceedings of the International Conference on Computing and Communication Systems (ICCCS)*, 2023, pp. 89–96.
- [14] K. Joshi and S. Singh, "Multilingual Text Mining for News Sentiment Analysis," *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, pp. 112–121, 2023.
- [15] T. Das and A. Roy, "Automated Categorization of News Articles with Emphasis on Political Bias Detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 1342–1354, 2023.