International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 2, Mar-Apr 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

NEWS FUSION: A MULTI-LINGUAL NEWS AGGREGATOR

*A.Sri Vaishnavi,** L. Sanjana, *** Sheetal Patri, **** Dr. K. Vaidehi
*Department of ADCE , Stanley College of Engineering and Technology for Women, Hyderabad
Email: vaishnaviachanta986@gmail.com
**Department of ADCE, Stanley College of Engineering and Technology for Women, Hyderabad
Email: sanjanalakkarsu@gmail.com
***Department of ADCE, Stanley College of Engineering and Technology for Women, Hyderabad
Email: patrisheetal26@gmail.com
****Department of ADCE, Stanley College of Engineering and Technology for Women, Hyderabad
Email: kvaidehi@stanley.edu.in

----------**********----------

Abstract:

The digital era has transformed the way news is generated, disseminated, and consumed globally. This review paper, explores advancements in news processing techniques, focusing on natural language processing (NLP) applications in multilingual resource management, genre classification, and sentiment analysis. By examining recent research, we aim to summarize developments in these areas, highlight the use of emerging algorithms, and identify current challenges and opportunities for future research. The reviewed studies provide a foundation for understanding the capabilities and limitations of current approaches, emphasizing the importance of adaptable, efficient systems in managing the rapidly evolving news ecosystem.

Keywords - NLP, Sentiment Analysis, Genre classification, Text to Speech Classification, Speech Synthesis Engine, Speech Recognition

----------**********----------

I. INTRODUCTION

The rapid expansion of digital news sources has transformed the landscape of information sharing, bringing new demands for efficient and adaptable processing techniques. As news content flows continuously from a vast range of sources and in numerous languages, there is a pressing need for sophisticated tools to analyze and classify this data accurately and in real-time. Traditional methods often fall short in handling the dynamic nature and sheer volume of news data, especially in an era where immediate insights and multilingual accessibility are essential. Natural language processing (NLP) has emerged as a powerful field offering solutions for analysing complex news content, with applications in sentiment analysis, genre classification, and multilingual management. This review, examines recent research in NLP techniques for news, exploring how innovations in this field are helping to meet modern challenges in information management. By surveying current methods, this paper highlights key advancements, compares approaches, and addresses limitations in processing digital news. As the demand for instant and precise information grows, further research into more efficient algorithms and better language models is crucial to improving news content analysis.

II. LITERATURE REVIEW

The field of text classification, especially within news genre categorization, resource management, and digital archiving, has made significant strides through the development of machine learning, computational linguistics, and interdisciplinary methodologies. An initial breakthrough in news genre classification involved using structure-based features to distinguish genres by examining linguistic and syntactic structures within articles, which has led to improved detection of diverse news types. However, the heavy reliance on extensive, genre-specific annotations for model accuracy limits the scalability of this approach, especially in rapidly evolving media landscapes [1]. In the context of multilingual resource management, significant advancements have been made through the use of metadata and classification techniques in

 ©IJSRED: All Rights are Reserved

library systems, which are designed to handle multilingual resources. These efforts enhance resource accessibility for users across linguistic backgrounds, yet they struggle with resource management at a larger, institutional level, highlighting the need for scalable models that can adapt to diverse datasets [2]. Another notable advancement in automated text analysis is the refinement of genre classification models tailored to specific genres. These models align text analysis techniques with the stylistic and structural features of particular genres, yielding more accurate classification in fields like the social sciences. However, this genre-specific focus often restricts the models' generalizability to other fields or document types, pointing to a gap in creating versatile, cross-domain models [3]. Further research has combined genre-revealing and subject-revealing features to refine text classification accuracy, integrating linguistic and contextual indicators. While effective, this approach requires detailed feature selection, which increases computational complexity, posing challenges for processing large datasets or real-time classification [4]. Online user profiling in question-answering forums, leveraging linguistic and behavioural cues, has provided personalized user experiences by identifying individual traits and preferences. This approach enables improved content recommendations but raises ethical concerns regarding user privacy, as extensive profiling may infringe upon personal data rights [5]. Studies on user comments in digital media have revealed that online commentary influences public perceptions and discourse around news topics, contributing to the framing of narratives. However, such studies often overlook regional or cultural variations in comment behaviour, suggesting a gap in understanding how different demographics engage with news content online [6]. In recent developments, transformer-based pre-trained models have redefined news summarization by significantly enhancing the precision and coherence of summaries. These models are highly effective for both extractive and abstractive summaries, yet their high computational requirements limit accessibility, particularly for resource-constrained environments [7]. BERT-based models have furthered genre classification accuracy, particularly for literary works, by addressing class imbalance and leveraging metrics like precision, recall, and F1-score. However, these models require advanced computational setups, such as multi-GPU environments, making them less accessible for smaller research institutions [8]. The analysis of comment evolution on platforms like Phox Smalltalk has revealed a taxonomy of class comment types, identifying inconsistencies in comment practices and highlighting a lack of standardization. While insightful, this research is limited by its platform-specific focus, which may not generalize to other software or environments [9]. Text segmentation models have enhanced our understanding of article structures, leading to improvements in multi-paragraph summarization by organizing content into coherent segments. However, such models face challenges when dealing with articles that lack clear thematic transitions, limiting segmentation effectiveness in these cases [10]. In genre-specific summarization, models capable of generating section-based heading summaries provide an organized flow of information across lengthy articles. This method enhances the readability of summaries but may struggle in cases where article structures are unconventional or lack well-defined sections [11]. In the domain of news summarization, the SEGNEWS dataset, annotated specifically for topic-focused themes, has demonstrated the effectiveness of using topic-specific prompts to improve summary relevance. This advancement addresses the need for fine-grained thematic analysis in news, but the dataset's specificity may limit its applicability to articles with more generalized topics or single-theme content [12]. The development of hybrid models combining BART and SVM has shown promise in news summarization by maintaining topic relevance while addressing summarization accuracy, making it suitable for articles with nuanced content. Nevertheless, the hybrid approach may require further refinement to handle topics that demand an extremely detailed understanding of context [13]. Recent advancements in sentiment analysis have leveraged deep learning techniques to improve the accuracy of opinion classification in textual data. The integration of recurrent neural networks (RNNs), convolutional neural networks (CNNs), and deep belief networks (DBNs) has enhanced sentiment classification, enabling more precise analysis of online reviews, social media posts, and public opinion [14]. However, challenges persist in the form of data sparsity, context dependence, and multilingual sentiment analysis, which require further exploration to improve cross-lingual sentiment detection. In the domain of speech synthesis, significant progress has been made through concatenative, statistical parametric, and hybrid text-to-speech (TTS) models. Concatenative synthesis, which relies on prerecorded speech units, has achieved high naturalness but lacks flexibility [15]. Text-to-speech (TTS) and speech-to-text (STT) systems have further evolved through the use of artificial neural networks (ANNs) and Hidden Markov Models (HMMs) to improve conversion accuracy. A hybrid ANN-HMM approach has

 ©IJSRED: All Rights are Reserved

demonstrated improvements in speech recognition accuracy, pronunciation modeling, and multilingual adaptation. However, high computational demands and limited availability of high-quality speech corpora for underrepresented languages pose challenges to widespread deployment [16]. Collectively, this body of research demonstrates substantial advancements in text classification, genre-based analysis, archival resource management, and sentiment analysis. However, persistent challenges remain in scalability, computational demand, model adaptability, and ethical considerations related to data privacy. The reliance on high computational resources, particularly for deep learning models like transformers, underscores the need for more efficient algorithms that can perform effectively within resource-constrained environments. Additionally, the varied contexts—ranging from news and social media to archives and libraries—highlight the importance of interdisciplinary research that bridges technical innovations with social and cultural understanding. Future work should focus on developing scalable models that integrate these diverse dimensions, enhancing the applicability and impact of automated classification, summarization, and archival techniques across domains. As the field advances, addressing these issues will be crucial for creating robust, adaptable systems capable of meeting the demands of increasingly complex and interconnected information landscapes.

III. METHODOLOGY

A. Data Collection and Preprocessing

The News Fusion platform aggregates news articles from multiple sources using web scraping and API integration. The collected data is categorized into predefined genres such as politics, stock markets, movies, and sports. Additionally, a multilingual filtering mechanism is implemented to personalize content delivery based on user-selected languages (English, Hindi, Telugu).

To ensure data consistency, preprocessing techniques such as stopword removal, tokenization, and normalization are applied. This improves the efficiency of subsequent sentiment analysis and speech processing tasks.

III.
IV.
V.
VI.

B. System Architecture and Frontend Development

The platform follows a modular architecture with a frontend-backend separation, developed using HTML, CSS, and JavaScript. The key modules include: News Categorization Module: Displays news articles under specific genres. Personalization Module: Recommends content based on user preferences. User Interaction Module: Implements commenting, note-taking, and saved headlines. Speech Processing Module: Enables speech recognition and text-to-speech functionalities.

C. Sentiment Analysis for News Classification

Sentiment analysis is implemented to classify news articles as Positive, Negative, or Neutral. The classification model utilizes natural language processing (NLP) techniques, including: Text Vectorization: Conversion of news text into numerical representations using TF-IDF (Term Frequency-Inverse Document Frequency). Sentiment Model: A pre-trained Naïve Bayes or Logistic Regression classifier is applied for text sentiment classification.

D. Speech Recognition for Hands-Free Interaction

A speech-to-text (STT) engine is integrated to allow users to interact with the platform through voice commands. The system utilizes: Web Speech API for real-time speech recognition. NLP-based Command Processing to extract keywords and perform appropriate actions (e.g., searching for news, navigating categories).

E. Text-to-Speech (TTS) for Enhanced Accessibility

To improve accessibility, a text-to-speech (TTS) engine is incorporated. The Web Speech API's Speech Synthesis is used to convert text-based articles into audio output, enabling users to listen to news.

F. Data Storage and User Preferences

The system maintains user interactions using localStorage, ensuring a seamless experience without requiring backend databases. The stored data includes: User Comments, saved News Articles with full content, personal Notes for future reference.

IV. CONCLUSION

In conclusion, we developed NewsFusion**,** an intelligent and multilingual news aggregation system that enhances the way users access, interact with, and consume news content. The system integrates news aggregation,

 ©IJSRED: All Rights are Reserved

sentiment analysis, speech recognition, and text-to-speech synthesis, creating a comprehensive platform tailored for diverse user preferences. By categorizing news into genres and incorporating language-based filtering, the platform ensures a personalized and user-friendly experience.

A key advancement in News Fusion is the sentiment analysis module, which provides users with a quick understanding of an article's tone—whether positive, negative, or neutral. This feature assists in filtering information based on sentiment, offering an additional layer of relevance to news consumption. Furthermore, speech recognition enables voice-based interactions, reducing dependency on manual input, while the text-to-speech (TTS) engine enhances accessibility for users who prefer audio-based news delivery. These features make News Fusion an inclusive and interactive news platform.

The system was designed with a lightweight, browser-based architecture, leveraging HTML, CSS, JavaScript, and local Storage for seamless functionality without requiring a backend database. Experimental evaluations confirmed the effectiveness of the sentiment classification model, the accuracy of speech recognition, and the efficiency of the text-to-speech system across different user conditions. Therefore the proposed system contributes to the evolving landscape of news aggregation platforms by addressing both user engagement and accessibility challenges, paving the way for next-generation intelligent news applications.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Stanley College of Engineering and Technology for Women for providing the essential resources, technical support, and encouragement throughout this research. The institution's commitment to academic excellence and innovation has been instrumental in enabling us to undertake and successfully complete this study.

We extend our heartfelt appreciation to Dr. K. Vaidehi Ma'am for her invaluable guidance, constructive feedback, and continuous support, which greatly contributed to the success of this study. Her expertise and mentorship have been a source of inspiration, helping us refine our approach and strengthen the quality of our research. We also acknowledge the unwavering support of our faculty members and peers, whose insightful discussions and suggestions have provided us with new perspectives and helped enhance the depth of our work. Their encouragement and willingness to share knowledge have played a crucial role in shaping the outcomes of this research.

Furthermore, we are deeply grateful to our families and friends for their constant motivation, patience, and belief in our capabilities. Their support has been our driving force, allowing us to remain dedicated and focused throughout this journey.

Finally, we would like to extend our appreciation to the broader research community, whose foundational work in news aggregation, sentiment analysis, speech recognition, and text-to-speech synthesis has served as a cornerstone for our study. This research would not have been possible without the collective contributions of scholars and developers in the field.

REFERENCES

[1] Dai, Z., Taneja, H., & Huang, R. (2018). Fine-grained Structure-based News Genre Categorization. Proceedings of the Workshop Events and Stories in the News 2018, 61-67. Association for Computational Linguistics.

[2] Mandal, S. (2018). Development of Multilingual Resource Management Mechanisms for Libraries. Journal of Library and Information Science, 44(2), 101-115.

[3] Ponciano, L. & Cardoso, P. J. S. (2021).Online Newspapers and the Comment Sections: Perception and Use. Asian Research Journal of Arts & Social Sciences, 15(1), 14-24.

[4] Sharoff, S., Wu, Z., & Market, K. (2010). Text genre classification with genre-revealing and subject revealing features. ResearchGate.

[5] Tjong Kim Sang, E. F., & De Rijke, M. (2013). Recommending tags with a model of human categorization

[6] Ziegele, M., Springer, N., Jost, P., & Wright, S. (2017). Online user comments across news and other content formats: Multidisciplinary perspectives, new directions. Studies in Communication and Media, 6(4), 315-332.

 ©IJSRED: All Rights are Reserved

[7] Gupta, A., Chugh, D., Anjum, A., & Katarya, R. (2021). Automated News Summarization Using Transformers. Journal of Information Science and Technology, 56(2), 108-125.

[8] Liu, S., Huang, Z., Li, Y., Sun, Z., Wu, J., & Zhang, H. (2020). DeepGenre: Deep Neural Networks for Genre Classification in Literary Works. Language Technologies Institute, Carnegie Mellon University.

[9] Rani, P., Panichella, S., Leuenberger, M., Ghafari, M., & Nierstrasz, O. (2021). What do class comments tell us? An investigation of comment evolution and practices in Pharo Smalltalk. Empirical Software Engineering, 26(112), 1-49

[10] Liu, Y., Zhu, C., & Zeng, M. (2021). End-to-End Segmentation-based News Summarization. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 16(4), 29-39.

[11] Finn, A., & Kushmerick, N. (2006). Learning to classify documents according to genre. Journal of the American Society for Information Science and Technology, 57(11), 1506-1518.

[12] Bahrainian, S. A., Feucht, S., & Eickhoff, C. (2022). NEWTS: A Corpus for News Topic-Focused Summarization. Proceedings of the Association for Computational Linguistics (ACL 2022), 493-503.

[13] Farrel Octavianus; Albert Wihardi; Muhamad Keenan Ario; Derwin Suhartono Automated Text Summarization and Topic Detection on News Aggregation System Using BART and SVM. IEEE International Conference on Big Data (Big Data), 994-1003.

[14] Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., & Rehman, A. (2017). Sentiment Analysis Using Deep Learning Techniques: A Review. *International Journal of Advanced Computer Science and Applications*, 8(6), 424-428.

[15] Khan, R. A., & Chitode, J. S. (2016). Concatenative Speech Synthesis: A Review. *International Journal of Computer Applications*, 136(3), 1-4.

[16] Nagdewani, S., & Jain, A. (2020). A Review on Methods for Speech-to-Text and Text-to-Speech Conversion. *International Research Journal of Engineering and Technology (IRJET)*, 7(5), 4459-4463.

 ©IJSRED: All Rights are Reserved