NEWS Aggregation System Leveraging NER and Classifiers: Summarization

Sharon Roshini Swaminathan and **Jayasri Suresh Vani**
Purdue University Fort Wayne
Fort Wayne, Indiana
swamsr01@pfw.edu, surej01@pfw.edu

Abstract

This study introduces a geography-aware news aggregation system that classifies news articles by region and generates contextual summaries. Using Named Entity Recognition (NER) for geographic identification and machine learning models—Logistic Regression (LR), Support Vector Machines (SVM), and XGBoost—for classification, the system organizes articles into predefined regions. These classified articles with respect to regions has summarization employed extractive techniques (TextRank, LexRank, BERTSum) and abstractive models (T5, BART, Pegasus), with extractive methods delivering higher ROUGE scores for region-specific content. The results achieved demonstrate the system's capability to enhance news aggregation through NER, classification, and summarization.

1 Introduction

With the explosion of digital news content, users are often overwhelmed by the volume of available information. Traditional news aggregation systems primarily focus on topics but often neglect the geographic context, a crucial factor in user preferences and relevance. This lack of geographic personalization results in fragmented or irrelevant news consumption experiences.

The primary objective of this project is to address this gap by building a news aggregation system that leverages Named Entity Recognition (NER) and machine learning classifiers to classify news articles by geographic region. Additionally, the system generates concise, region-specific summaries to enhance coherence and user experience. This approach combines advanced classification techniques with both extractive and abstractive summarization methods to provide users with personalized and contextually relevant news.

The project is motivated by the increasing demand for personalized news delivery. Current systems fail to address regional relevance, which is vital for users seeking location-specific information. By integrating classification and summarization techniques, this system aims to provide precise, tailored information, enabling faster decision-making and improving the overall news consumption experience.

This system classifies news articles into predefined geographic regions using NER and machine learning classifiers. It generates high-quality summaries through both extractive (TextRank, LexRank, BERTSum) and abstractive (T5, Pegasus, BART) techniques. The performance of classifiers and summarization models is evaluated using metrics such as accuracy, ROUGE, and BLEU to ensure concise, region-specific, and coherent information delivery.

The remainder of this paper is organized as follows: Section 2 reviews related work on NER, classification, and summarization techniques. Section 3 details the dataset and preprocessing strategies. Section 4 presents results, analysis, and evaluation metrics. Section 5 concludes with insights, challenges faced, and Section 6 narrates the directions for future work.

2 Related Works

Named Entity Recognition (NER) plays a crucial role in natural language processing, and numerous tools have been developed to address this task. A comparative study of seven widely used NER libraries, including Stanford NER, spaCy, NLTK, Polyglot, Flair, GATE, and DeepPavlov, highlights their varying performance across languages and datasets (Vychegzhanin and Kotelnikov, 2019). The study found DeepPavlov excelling in Russian and Flair performing best in English, providing valuable insights for selecting appropriate tools for specific tasks.

In the domain of text summarization,

transformer-based models like GPT, T5, BART, and PEGASUS have demonstrated their effectiveness, with T5 outperforming others in tasks requiring concise and informative summaries (Dharrao, 2023), (Asmitha et al., 2024). Further studies show that integrating BART with TextRank enhances thematic alignment in news summaries (Chen and Song, 2021). Comparative analyses of TextRank and LexRank reveal their respective strengths in extractive summarization, with TextRank excelling in thematic extraction (Ghorpade et al., 2024).

A comprehensive review of summarization methods highlights the trade-offs between computational efficiency and summarization quality, with transformer models like T5 excelling in high-quality abstractive summaries (Steven Shearing). Similarly, T5's versatility in retaining contextual meaning in generated summaries has been emphasized (Prof. Kalyani Pendke, 2023). An ensemble approach combining Logistic Regression, SVM, and XGBoost has demonstrated improved performance in extractive summarization tasks (Singh et al., 2020). These advancements in NER and summarization techniques provide a solid foundation for developing robust systems capable of processing and summarizing large volumes of textual data efficiently.

3 Methodology

This section outlines the methodology adopted for integrating Named Entity Recognition (NER) in both extractive and abstractive News summarization. The code for this project is maintained on GitHub. [1] The methodology is structured into several key phases:

3.1 Proposed Architecture

This system architecture, as shown in Figure 1, is designed to process raw news articles, classify them by geographic regions, and generate high-quality summaries. The workflow begins with preprocessing, where Named Entity Recognition (NER) extracts key geographic entities, and the articles are mapped to predefined regions. The text is then vectorized using TF-IDF to create numerical representations for classification. Logistic Regression, Support Vector Machines (SVM), and XGBoost are employed for region classification, with XGBoost delivering the best performance due to its ability to handle complex interactions and sparse data. The classification output is independent of the summarization process, ensuring modularity.

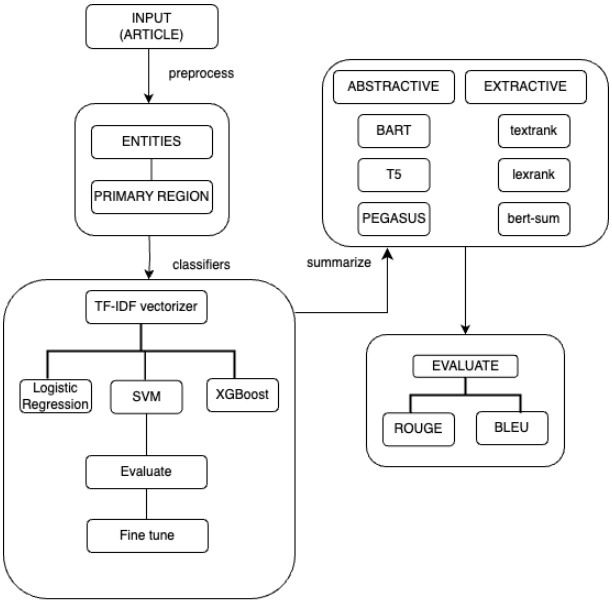

Figure 1: Comparison of classification performance across Logistic Regression, SVM, and XGBoost.

Summarization is performed using two approaches: extractive methods (TextRank, LexRank, BERT-Sum) and abstractive methods (T5, BART, Pegasus). Extractive methods directly select relevant sentences, while abstractive methods generate paraphrased summaries. Evaluation consolidates the performance of both tasks, using metrics such as accuracy for classification and ROUGE/BLEU for summarization. This system effectively combines NER, classification, and advanced summarization techniques to deliver concise, region-specific, and high-quality news summaries tailored to user preferences.

3.2 Data Collection

The dataset used for this project is the CNN/DailyMail dataset hosted on Hugging Face (`abisee/cnn_dailymail`). This English-language dataset contains over 300,000 unique news articles authored by journalists from CNN and the Daily Mail. It supports both extractive and abstractive summarization tasks.

Each data instance consists of three main fields: *id*, *article*, and *highlights*. The *id* is a hexadecimal SHA1 hash of the article's source URL, while the *article* provides the full text, and the *highlights* offer a concise summary written by the author. The

[1] https://github.com/Jayasri2021/News-Summarization-NLP

mean token count for articles is 781, and for highlights, it is 56. The dataset is structured into three splits: **train**, **validation**, and **test**, with 287,113, 13,368, and 11,490 instances respectively.

3.3 Preprocessing

The implemented steps on cleaning and standardizing the dataset to ensure preprocessing are :

First, tokenization was applied to split the text into individual tokens, and all characters were converted to lowercase to eliminate variability caused by case sensitivity. Non-informative elements such as stopwords (e.g., "and", "the") and punctuation were removed to retain only meaningful content. Additionally, URLs were stripped from the text to eliminate irrelevant and non-contextual information that could interfere with analysis. The preprocessing pipeline utilized tools like ‘spaCy‘ for tokenization and stopword filtering, combined with regular expressions for URL removal. To ensure flexibility and data integrity, preprocessed articles were stored as new columns in the dataset, allowing the original data to remain intact for future reference.

3.4 Named Entity Recognition

Named Entity Recognition (NER) was essential for categorizing news articles by geographic regions, ensuring region-specific content delivery. Using spaCy's pre-trained pipeline, geopolitical entities (GPE) were efficiently identified. Recognized locations were then mapped to predefined geographic regions, including North America (NA), Europe (EU), Asia (ASIA), and others, through a custom dictionary. This mapping ensured consistent and accurate regional classification.

3.5 Classifier Models

After preprocessing and NER-based region mapping, the data is vectorized using TF-IDF, which highlights region-specific terms by downweighting common words and emphasizing rare but significant ones. This aligns with the goal of precise geographic categorization.

The TF-IDF vectors serve as input for classifiers like Logistic Regression (LR), SVM, and XGBoost. Among these, XGBoost demonstrates the best performance, effectively handling sparse, high-dimensional data and capturing complex interactions, resulting in higher classification accuracy.

The output predictions of these classifiers are used as inputs for the subsequent summarization task to ensures that summaries are contextually relevant to the geographic categories.

3.6 Implementation

The three classifiers—Logistic Regression (LR), Support Vector Machines (SVM), and XGBoost—greatly enhanced the mapping of entities (primary regions) beyond the capabilities of Named Entity Recognition (NER). While NER efficiently identified geopolitical entities, these classifiers leveraged the structured representations from TF-IDF vectorization to accurately categorize articles into predefined regions. Logistic Regression served as the baseline model, providing a straightforward and interpretable starting point. It effectively used TF-IDF features to capture term importance relative to the dataset and achieved moderate classification accuracy. The baseline hyperparameters (`C=1`, `penalty=l2`, and `solver=liblinear`) balanced model complexity and regularization, demonstrating the strength of LR in handling sparse data while offering a point of comparison for more complex models.

SVM improved upon Logistic Regression by leveraging its ability to find an optimal hyperplane in high-dimensional spaces, which is particularly suitable for the sparse TF-IDF vectors. Using a linear kernel, SVM achieved a higher baseline accuracy, demonstrating superior performance in handling class separability. Hyperparameter tuning with `RandomizedSearchCV` focused on optimizing regularization strength (`C`), penalty type, and iteration limits. While the tuned hyperparameters (`C=1`, `penalty=l2`, `max_iter=2000`) maintained similar accuracy, the model effectively handled imbalanced and sparse datasets, highlighting the trade-offs between simplicity and performance in high-dimensional classification.

XGBoost outperformed both LR and SVM by efficiently modeling complex feature interactions through its boosting framework. Its baseline accuracy was achieved by leveraging TF-IDF's sparse matrix representations and encoding labels to ensure compatibility with the classifier. Hyperparameter tuning with `GridSearchCV` focused on parameters like learning rate (e.g., 0.01), maximum tree depth (e.g., 5), and column sampling ratios. These optimizations allowed XGBoost to balance complexity and generalization, significantly improving performance on underrepresented classes. Its iterative boosting process ensured the correction of weak predictions, making it robust against overfit-

ting and capable of extracting nuanced interactions from the data.

3.7 Summarization

The output of the fine-tuned XGBoost classifier, which categorizes news articles into geographic regions, serves as the input for all summarization tasks. This integration ensures that the summarization process is tailored to the regional classification.

3.8 Extractive summarization

For extractive summarization, TextRank, LexRank, and BERTSum were employed to create concise summaries. TextRank, a graph-based algorithm, ranks sentences by importance using cosine similarity and PageRank, effectively extracting key region-focused content without the need for labeled data. LexRank builds on this by adding a cosine similarity threshold, improving precision in sentence selection and producing coherent summaries aligned with regional themes. BERTSum, leveraging pre-trained BERT embeddings, outperformed the other methods by capturing rich semantic context, ranking sentences based on relevance, and delivering contextually accurate summaries.

3.9 Abstractive summarization :

For abstractive summarization, BART, T5, and PEGASUS were implemented to generate fluent and rephrased summaries. BART, fine-tuned on 1024-token inputs with beam search, excelled in creating coherent summaries for lengthy and complex articles. T5, with its versatile text-to-text framework and tokenized inputs capped at 512 tokens, efficiently produced concise and contextually relevant summaries by balancing brevity and informativeness. PEGASUS, pre-trained with a gap-sentence generation objective, focused on predicting key missing sentences, generating concise and detail-oriented summaries optimized for the dataset. Together, these models produced high-quality, region-specific summaries for news articles.

4 Results and Analysis

All the results of the experiments to classify and summarize the news articles are infered in this section:

The results from Table1 demonstrate that XGBoost performs the best in accurately identifying these regions, achieving the highest accuracy of 79 percent. This indicates its robustness in handling the classification task and its ability to model complex relationships within the dataset. SVM follows closely with 0.78 accuracy, showing that it is also effective but slightly less consistent than XGBoost. Logistic Regression, with an accuracy of 0.74, struggles to match the performance of the other two models, potentially due to its linear nature, which might not fully capture the nuanced relationships within the data.

Metric	Logistic Reg.	SVM	XGBoost
Accuracy	74%	78%	79%
W. Precision	78%	77%	73%
W. Recall	75%	68%	73%
W. F1-Score	71%	60%	69%

Table 1: Classification metrics comparison for Logistic Regression, SVM, and XGBoost.

Before fine-tuning, XGBoost and Logistic Regression demonstrate a stronger capability in capturing the primary region's class distribution, as evidenced by higher recall values. However, the weighted F1-Score for SVM (0.60) suggests that it initially struggles with balancing the precision and recall across all regional classes, likely leading to misclassifications for less represented regions.

Metric	Logistic Reg.	SVM	XGBoost
Accuracy	74%	78%	79%
W. Precision	78%	77%	73%
W. Recall	75%	68%	73%
W. F1-Score	71%	60%	69%

Table 2: Classification metrics comparison for Logistic Regression, SVM, and XGBoost.

Metric / Parameter	Logistic Regression (LR)	SVM	XGBoost (XGB)
Metrics			
Accuracy (%)	74	81	83
Precision (Macro Avg)	90	82	85
Recall (Macro Avg)	88	81	84
F1-Score (Macro Avg)	71	73	77
Tuned Parameters			
Regularization (C)	1.0	1.0	1.0
Penalty	L2	L2	L2
Learning Rate	0.001	0.01	0.001
Epochs	10	10	5

Table 3: Comparison of Classification Metrics and Hyperparameters for Logistic Regression, SVM, and XGBoost after fine-tuning.

Table 2 , After fine-tuning, the second table highlights significant improvements in the models' ability to classify primary regions accurately. XGBoost

achieves the highest accuracy (0.83) and macro-average metrics, including precision (0.85) and recall (0.84), indicating its strength in correctly identifying regional classes, even for less represented ones. For example, regions such as the Middle East (ME) and South America (SA), which might be more challenging due to limited data, are more effectively classified after fine-tuning, as evidenced by higher F1-Scores. SVM also shows substantial improvement, with an accuracy of 0.81 and a macro-average F1-Score of 0.73, making it a competitive model for region classification.

The classified models are saved and loaded for summarizing the articles. From Table 3 , it is obvious that Extractive models outperform abstractive models across all metrics, with BERTSum emerging as the top performer. It achieves the highest ROUGE-1 (0.80), ROUGE-2 (0.81), ROUGE-L (0.83), and BLEU (0.51) scores, indicating its strength in accurately capturing and preserving key information while maintaining the structure of the original text. TextRank and LexRank follow closely, proving their effectiveness in generating concise and relevant summaries.

	Abstractive			Extractive		
Metric	**T5**	**BART**	**PEGASUS**	**LexRank**	**TextRank**	**BERTSum**
ROUGE-1 (F1)	0.32	0.42	0.28	0.76	0.81	0.80
ROUGE-2 (F1)	0.20	0.32	0.19	0.71	0.74	0.81
ROUGE-L (F1)	0.32	0.424	0.16	0.76	0.81	0.83
BLEU	0.39	0.43	0.27	0.39	0.37	0.51

Table 4: Performance comparison of abstractive and extractive summarization models.

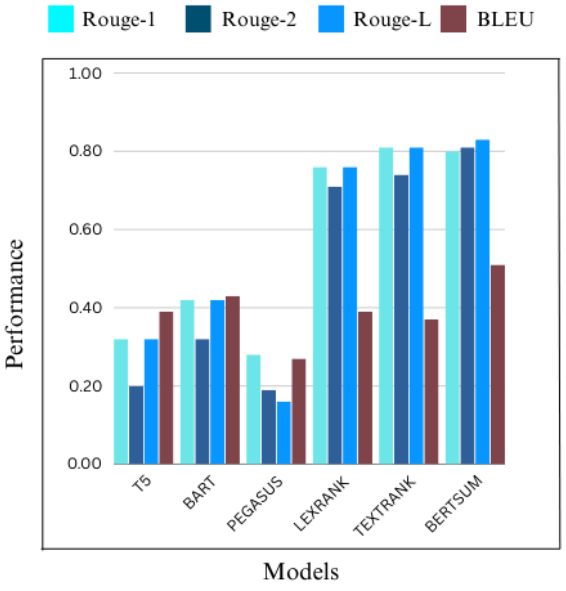

Figure 2: ROUGE and BLEU score for T5, BART, PEGASUS, LexRank, TextRank, BERTSum

From figure 2 , it is inferred that among abstractive models, BART delivers the best results with ROUGE-1 (0.42), ROUGE-2 (0.32), ROUGE-L (0.424), and BLEU (0.43). Its ability to produce coherent and human-like summaries makes it a strong choice for tasks requiring fluency and paraphrasing. T5 provides moderate performance, lacking the coherence and precision of BART. PEGASUS, however, struggles significantly, with the lowest scores across all metrics, indicating challenges in producing fluent and contextually accurate summaries.

5 Conclusion

In conclusion, the classification of regions played a pivotal role in ensuring the relevance and coherence of the summaries. By accurately categorizing articles into their primary regions using advanced classifiers like XGBoost, the system ensured that the summarization models operated on contextually appropriate and region-specific content. XGBoost's superior performance in classification enhanced the quality of input for the summarization task, directly influencing the generation of summaries tailored to geographic contexts.

For summarization, extractive methods outperformed abstractive models, with BERTSum emerging as the most effective choice. It preserved critical information and maintained the structure of the original text, making it highly suitable for delivering regionally relevant content. However, for tasks requiring human-like summaries with paraphrased text, BART demonstrated its strength as the best abstractive model, producing fluent and coherent summaries. While T5 offered moderate performance, PEGASUS required further optimization to match the effectiveness of other models. Hence, the combination of accurate regional classification and robust summarization techniques resulted in a system capable of delivering concise, relevant, and geographically tailored news summaries.

6 Future Works

In the future, the system can be improved by adding multi-label classification, which would allow a single article to belong to multiple regions. This would make the classification more accurate for articles that cover topics across different regions. Another area for improvement is reducing errors in extractive summarization, such as the inclusion of irrelevant or disconnected content. Additionally,

grouping articles by topics within each region can help create summaries that are more focused and aligned with specific themes.

References

M Asmitha, C.R. Kavitha, and D Radha. 2024. Summarizing news: Unleashing the power of bart, gpt-2, t5, and pegasus models in text summarization. In *2024 4th International Conference on Intelligent Technologies (CONIT)*, pages 1–6.

Yisong Chen and Qing Song. 2021. News text summarization method based on bart-textrank model. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 2005–2010.

Mishra M. Kazi A. Pangavhane M. Pise P. Bongale A.M. Dharrao, D. 2023. Summarizing business news evaluating bart, t5, and pegasus for effective information extraction. *Revue d'Intelligence Artificielle*, 38(3):847–855.

Shreyas Ghorpade, Ayesha Khan, Akhelesh Chaurasia, Vir Rao, and Aditi Chhabria. 2024. A comparative analysis of textrank and lexrank algorithms using text summarization. In *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 379–393, Singapore. Springer Nature Singapore.

Tanuja Kadwe Bhavik Balpande Chitrani Somkuwar Utkarsh Zode Prof. Kalyani Pendke, Achal Lute. 2023. Designing a system using nlp for summarizing a text and retaining crucial points in a text. *JETIR*, 10:a503–a508.

Prabhjot Singh, Prateek Chhikara, and Jasmeet Singh. 2020. An ensemble approach for extractive text summarization. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–7.

Benjamin Wellner Liz Merkhofer Steven Shearing, Abigail Gertner. Automated text summarization: A review.

Sergey Vychegzhanin and Evgeny Kotelnikov. 2019. Comparison of named entity recognition tools applied to news articles. In *2019 Ivannikov Ispras Open Conference (ISPRAS)*, pages 72–77.