From Chaos to Clarity: Enhancing Personalized News Consumption in the Digital Age

1st Aiman Gohar
dept. FCSE Artificial Intelligence, GIKI
Swabi, Pakistan
u2020059@giki.edu.pk

2nd Dr Raja Hashim Ali
Associate Professor FCSE, GIKI
Swabi, Pakistan
hashim.ali@giki.edu.pk

***Abstract*—Amidst the proliferation of online news platforms, individuals face a daunting challenge in navigating the vast sea of information to stay informed about global events. The shift from traditional newspapers to digital news sources necessitates efficient methods for personalized information consumption. This study addresses the contemporary need for effective news classification and summarization, acknowledging the overwhelming nature of online news and the time constraints faced by readers. While existing solutions often focus on either news classification or summarization, this research uniquely endeavors to optimize both tasks, presenting a comprehensive approach that encompasses a diverse array of machine learning models, including SVM, LR, MNB, RF, and DTC, as well as exploring various word embedding techniques like GloVe, word2vec, CBOW, Skip-gram, TF-IDF, and Fasttext. Extractive summarization techniques, Textrank, and LexRank are further scrutinized to analyze the working and identify the most effective method for condensing news articles. This study utilizes three extensive datasets, namely CNN News Articles from 2011 to 2022 [1], the News Article Category dataset [2], and the BBC News Summary dataset [3]. Support Vector Machine (SVM) and logistic regression (LR) using the TF-IDF word embedding technique achieved remarkable success, with precision and recall values of 0.88%. The results not only provide insights into the optimal combinations of algorithms for news classification but also offer a refined understanding of extractive summarization techniques. By addressing this dual challenge, the study contributes a valuable methodology that enhances information accessibility and personalization in the fast-paced realm of online news consumption.**

***Index Terms*—News Classification, News Summarization,Personalized information consumption,Word Embedding Techniques, Machine Learning, Extractive Summarization**

I. Introduction

The internet has ushered in a significant revolution worldwide, simplifying information retrieval and making it more convenient. Individuals sitting in distant parts of the world can now quickly learn about various global accidents or events through social media platforms or other online news platforms. Currently, most news and media organizations maintain their websites and applications to disseminate information about events to a broad audience, in addition to traditional newspapers. People now prefer online news reading over newspapers because almost everyone in the present world has internet access, enabling them to learn about an event moments after it occurs [4]. In contrast, acquiring information through newspapers is a time-consuming process. According to a survey conducted by the Pew Research Center [5], individuals in the United States aged 18 to 29 show a greater inclination to receive news indirectly through social media than directly from print newspapers or news websites.

A substantial volume of data is uploaded to the internet on a daily basis. The increasing prevalence of news on social media provides individuals with more opportunities to encounter news incidentally while engaging in online activities. In their 2021 research [6], Park and Kaye investigated the impact of incidental exposure to news through social media on perceptions of news accessibility, efficacy, and actual consumption. The findings from a two-wave survey of South Korean adults revealed a negative association between incidental exposure and news consumption across traditional and online media. Additionally, there was a reinforcing effect on the perception that news finds the individual, and the perception that news finds the individual played a mediating role in the relationship between incidental exposure and news consumption on both traditional and online platforms. Besides social media platforms, numerous online news platforms are available, keeping people informed about news across various fields such as sports, entertainment, and more. This study utilizes three extensive datasets, namely CNN News Articles from 2011 to 2022, the News Article Category dataset, and the BBC News Summary dataset to perform news categorization and news summarization tasks.

News platforms aim to upload events as fast as possible on their online platforms so that people can acquire knowledge about them promptly. However, the process of categorizing news articles under the correct category is a manual one, requiring a person to read the content and determine the field or class to which the article belongs. Given the volume of news across multiple online platforms, it becomes impossible for readers to access all the rich information due to time constraints. People also prefer to read news according to their preferences. To address these issues, this paper proposes an approach to categorize news into different classes based on the content present in the news article and to summarize news articles using extractive summarization techniques. This allows readers to acquire important information according to their preferences in less time.

978-1-6654-5992-1/22/$31.00 ©2022 IEEE

With advancements in machine learning, deep learning, and natural language processing, document categorization and summarization have become achievable tasks. Numerous research works have been conducted to tackle the task of news categorization and summarization using advanced machine learning and deep learning techniques. Research extends beyond the English language to categorize news in various languages worldwide, such as Urdu [7], Turkish [8], Indonesian [9], and Chinese [10], employing different word embedding techniques, machine learning classifiers, and deep learning techniques. Various research works have performed summarization of news documents to provide important information in fewer sentences, giving readers the opportunity to acquire as much knowledge as possible. Two text summarization approaches, extractive and abstractive summarization, are generally used. Extractive summarization algorithms maintain content originality by extracting key sentences or phrases from the text without changing them. Techniques like LexRank [11], Latent Sentiment Analysis (LSA) [11], and Textrank [12] have been explored. Abstractive summarization algorithms, on the other hand, do not maintain content originality but generate a summary by paraphrasing or creating new sentences. Techniques like Pegasus [11], BART [11], and seq2seq models [13] have been explored.

This paper extends the groundwork established by previous research efforts and aims to advance the boundaries of news classification and categorization tasks. It does so by exploring various machine learning classifiers, word embedding techniques, and extractive summarization algorithms. The objective is to categorize unstructured documents and condense lengthy news articles. This facilitates customized news consumption, enabling readers to access information according to their preferences and obtain a substantial amount of information in a shorter period.

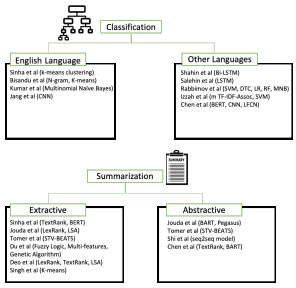

Fig. 1. Categorization of Approaches with Exemplary Papers

II. Literature Review

The literature review presents a comprehensive overview of text processing, classification, summarization, and clustering in the context of news articles, blogs, and diverse languages. Document clustering techniques [14] [15] are explored, demonstrating the effectiveness of text embedding methods like TF-IDF and Word2Vec [16]. Machine learning algorithms, including SVM and Naïve Bayes [17], prove valuable for news categorization, showcasing superior results with specific approaches. Studies in various languages, such as Bangla [18] [19], Uzbek [20],chinese [21], and indonesian [22] language , highlight the efficacy of diverse machine-learning models in news classification. Significant advancements in summarization include the comparison of extractive and abstractive techniques [11], with abstractive methods showing superiority. The integration of extractive and abstractive summarization techniques [23] contributes to further advancements. An automatic summarization model based on fuzzy logic rules [24] , multi-feature, and Genetic Algorithm outperforms existing methods. A comprehensive survey of neural abstractive text summarization [13]emphasizes techniques for enhanced readability. Summarization's significance extends to the blogosphere [12], with various algorithms explored for blog summarization. Multi-document summarization challenges are addressed [25], identifying unifiable news articles based on trending topics. An improved strategy for abstractive summarization, combining TextRank and BART [26], showcases performance enhancements over single BART models. This literature review 1 underscores the dynamic landscape of classification and summarization techniques across diverse domains and languages.

In today's world, we encounter a massive number of unstructured news documents. It has become a major and important task to make these documents readable and categorize them for easy extraction of information. Machine learning algorithms have demonstrated remarkable results in categorization tasks, making them the preferred choice for many research works. Document clustering is an unsupervised machine-learning technique that achieves the classification task by creating clusters or groups of related documents. Sinha et al [14] proposed the best method to cluster documents into groups that have the same categories or comparable contexts. BBC News dataset having articles of five different categories namely entertainment, sports, business, politics, and technology is used by generating their summary using BERT and text rank summarization algorithms. Different embedding techniques: TF-IDF, Word2Vec, Doc2vec, and GloVe along with dimensionality reduction were applied to the summarized articles using different vector lengths and they are clustered later using k-means clustering, observations regarding which stack of algorithms preserves the semanticity and context of the given documents belonging to each category were made. TF-IDF and Word2vec embedding techniques were found to give the best result in each news article category. Bisandu et al [15] proposed a new clustering technique using N-gram, K-means clustering, and improved sqrt-cosine similarity measure as a similarity measure. Detailed experiments were conducted using Reuters and 20Newsgroups datasets, and they concluded that the proposed clustering technique outperforms the existing clustering techniques. Kumar et al [17] performed a current affairs news categorization task using IDF and a multinomial Naïve Bayes machine learning algorithm. The study was

performed on 1800 news articles collected from three sources namely The Hindu, NDTV, and Indian Express website. News articles belonging to eight different categories namely business and economy, education and career, entertainment, food and health, international, politics and governance, science and technology, and sports, and found that the proposed model performs much better in news classification tasks when compared to other machine learning algorithms: Logistics Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbours (KNN) and Random Forest (RF). Different evaluation matrices: confusion matrix, accuracy, recall, and F1-score are used to evaluate the performance of the proposed model.

Deep learning algorithms, such as Convolutional Neural Networks, have demonstrated remarkable results in various classification tasks. Jang et al [16] conducted news and tweet classification using the CNN approach. They experimented with different word embedding techniques—word2vec, CBOW (Continuous Bag of Words), and Skip-gram—and discovered that word2vec-based Convolutional Neural Networks (CNNs) achieved better classification accuracy compared to CNNs with CBOW and Skip-gram.

In our diverse world, where people speak various languages and belong to different communities and countries, research has been conducted in languages like Hindi, Arabic, Indonesian, German, etc. News classification, a crucial task, has been explored in different languages. Shahin et al [18] conducted the classification of Bangle news headlines. They collected 100k news articles from the renowned news platforms of Bangladesh spanning eight different categories and used Bangla stop words for the removal of stop words and to achieve better results. For the vectorization of the text Genism and fastText models were used and hence the text was made compatible to be used as input for the three machine learning models to perform classification. Among the three machine learning models, Bi-LSTM performed better with an accuracy of 85.14%. Similarly, Salehin et al [19] performed Bangla news classification using five machine learning classifiers and two neural network algorithms and from their comparison concluded that LSTM give better results for the news categorization task. In a separate study, Rabbimov et al [20] conducted news classification by collecting news articles spanning 10 different categories written in Uzbek language from the ”Daryo” online news platform. They utilized TF-IDF and N-gram techniques for feature extraction, employing five machine learning models: Support Vector Machines (SVM), Decision Tree Classifier (DTC), Random Forest (RF), Logistic Regression (LR), and Multinomial Naïve Bayes (MNB). Using 5-fold cross-validation as a hyperparameter for text classification, they achieved a remarkable accuracy of 86.88%. Izzah et al [22] focused on enhancing text classification through an improved term weighting scheme, mTF-IDF-Assoc. Building on the previous TF-Assoc approach, mTF-IDF-Assoc incorporates the document length (DL) to optimize term weighting distribution in multiclass classification. Utilizing an SVM classifier and 10-fold cross-validation, the proposed scheme surpasses TF-IDF, TF-ICF, and TF-Assoc, achieving an average accuracy of 82.322%, demonstrating its efficacy in improving text categorization performance. This study utilized a dataset sourced from the Twitter account of the private radio station PR FM (@PRFMnews) in Bandung, comprising short texts in Indonesian. Classifying long texts is a challenging task. Chen et al [21] conducted long text classification in Chinese. This paper proposes an LFCN architecture, leveraging BERT and CNN, to tackle the challenge of long text classification in Chinese news. The suggested framework aims to capture both local and global text features. The data were collected from various Chinese news platforms in China, referred to as the MCNews dataset, and a LEAD-based extractive summarization algorithm named DLn was employed to extract key sentences or information for the classification task. Thorough experiments were performed on the THUCNews and MCNews datasets, revealing that the novel approach is both reasonable and effective.

Digital technology has significantly enhanced the accessibility of news from various categories and regions worldwide. However, with the increase in the number of online news platforms, it has become a challenging task for readers to read all the articles and absorb all the information due to their busy lives. Text summarization algorithms are crucial for addressing this issue, allowing readers to quickly skim through numerous articles and glean essential information in less time and with fewer sentences. Two text summarization approaches extractive and abstractive summarization are generally used. Extractive summarization algorithms maintain the content originality by extracting the key sentences or phrases from the text without changing them and thus the summary contains the important sentences of the original text. The abstractive summarization algorithms do not maintain the content originality but generate the summary of the text by paraphrasing or making new sentences that convey all the important information present in the text.

A comparison between the mentioned two summarization approaches is done in the study performed by Jouda et al [11]. They used short articles belonging to the sports and politics categories of the CNN/DM news dataset and performed a comparison by using LexRank, and LSA extractive summarization algorithms and two abstractive algorithms: BART and Pegasus. A web application was developed to give users a platform to summarize a text and perform comparisons between summarization techniques. ROUGE and BLEU are used as evaluation matrices and 10 participants were encouraged to evaluate the performance of the summarization algorithms using the summaries of 20 sports articles generated by these algorithms. Abstractive summarization outperformed according to the results of the evaluation matrices while in human evaluation extractive summarization was found to give the best result in the summarization task.

In a separate study, Tomer et al [23] addressed the limitations of abstractive and extractive summarizations by integrating these two approaches through a novel method called STV-BEATS. Key sentences from the news document are extracted using extractive summarization with skip-thought vectors and

a deep autoencoder. These extracted sentences are then inputted into a GRU-based bi-encoder and decoder to generate rephrased or new sentences through abstractive summarization, providing a summary of the news document. Tomer et al.'s study contributes to advancing the field of text summarization by introducing an integrated framework that maximizes the advantages of extractive and abstractive techniques.

Du et al [24] introduced a novel automatic summarization model for news text based on fuzzy logic rules, multi-feature, and Genetic Algorithm (GA). The model evaluates word and sentence features, scoring each word based on importance and combining sentence features using a linear combination with weights determined by GA. Special news elements such as time, place, and characters are extracted as keywords. The final score is calculated using a fuzzy logic system to achieve automatic summarization. Comparative results with other methods, including Msword, System19, System21, System 31, SDS-NNGA, GCD, SOM, and Ranking SVM, using the ROUGE assessment method on the DUC2002 dataset, demonstrate the proposed method's superiority over the mentioned approaches.

Shi et al [13] offered a comprehensive survey of neural abstractive text summarization using sequence-to-sequence (seq2seq) models, focusing on techniques to enhance saliency, fluency, and human readability in generating high-quality summaries. The survey categorized techniques into network structures, training strategies, and summary generation algorithms, with attention to efficiency and parallelism during model training. It also briefly reviewed models initially proposed for language modeling and later applied to summarization. The authors introduce an open-source toolkit, Neural Abstractive Text Summarizer (NATS), and conduct extensive experiments on the CNN/Daily Mail dataset to assess the effectiveness of various neural network components. The study concludes by benchmarking two NATS-implemented models on two datasets, Newsroom and Bytecup.

Summarization is crucial for distilling relevant information, especially in the context of the vast blogosphere. Deo et al [12] investigated the suitability of three text summarization algorithms: Textrank, LexRank, and Latent Sentiment Analysis (LSA) for blogs, aiming to identify the most accurate tool for trustworthy automatic summarization. The study emphasized the significance of summarizing blogs, which play a pivotal role in expressing public sentiment and hold value for various fields like psychology, anthropology, economics, and political science. The research emphasized the need for algorithms tailored to diverse content and computational principles for effective blog summarization. In a separate study, Singh et al [25] proposed work focused on multi-document summarization by identifying unifiable news articles related to top trending topics/hashtags. A framework was introduced to cluster news articles based on trending topics and capture opinion diversity. The study explored the effectiveness of text vectorization methods, including bag-of-words with tf-idf scores, word embeddings, and document embeddings. Text clustering using k-means is applied to news articles from the DUC 2004 benchmark dataset. The comparative analysis evaluates the performance of different vectorization methods in terms of purity, aiming to enhance the summarization process.

Chen et al [26] addressed the challenge of topic deviation in abstractive summarization by proposing an enhanced strategy. Leveraging TextRank and the BART model, the method involved initially using TextRank and BART to extract and generate summaries from news text. The results from both methods were then combined, emphasizing key sentences to enhance thematic coherence. The final summarization was obtained by inputting the improved texts into the BART model again. Experimental results demonstrate that this approach outperforms a single BART model, yielding improvements in Rouge-1, Rouge-2, and Rouge-L recall scores by 1.5%, 0.5%, and 1.3%, respectively.

A table summarizing the characteristics of the existing literature along with the novelty of the proposed work is shown (as shown in Table I).

III. Our Contribution

A. Gap Analysis

In the expansive landscape of research on news classification and summarization, there exists a discernible gap in the exploration of a comprehensive and integrated methodology applied to a substantial dataset representative of news articles from a reputable source. While previous studies have delved into diverse aspects of text processing, clustering, and classification across various languages and domains, the specific focus on optimizing both news classification and summarization tasks concurrently, using an extensive datasets like the CNN News Articles spanning over a decade, remains notably unexplored. The majority of existing works tend to concentrate on either classification or summarization, often employing smaller datasets or narrowly defined scopes. Furthermore, the proposed study introduces a varied set of machine learning models and word embedding techniques, encompassing SVM, LR, MNB, RF, DTC, GloVe, word2vec, CBOW, Skip-gram, TF-IDF, and Fasttext, thereby extending beyond the conventional subset of methodologies commonly employed in prior research. This novel approach not only addresses the practical challenges faced by readers in the digital era but also bridges the gap in the literature by providing a nuanced understanding of optimal algorithmic combinations for news classification and a refined insight into extractive summarization techniques. The utilization of the CNN News Articles from 2011 to 2022 [1], the News Article Category [2] , and the BBC News Summary [3] datasets enhances the robustness and relevance of the findings, contributing to the advancement of knowledge in the field and underscoring the need for a holistic approach in tackling the complexities of news processing and consumption.

B. Research Questions

In this research endeavour, our primary objective is to address the following research questions.

1) RQ1: How can a diverse range of machine learning models, including SVM, LR, MNB, RF, and DTC, be effectively employed to optimize news classification in

TABLE I
LITERATURE REVIEW TABLE SHOWING THE CONTRIBUTIONS OF VARIOUS AUTHORS FOR QUANTIZATION OF NETWORKS.

Year	Paper Author	Method(s)	Dataset	Contribution(s)	Drawback / Limitations	Categorize	Summarize	Accuracy	Evaluation Metrices
2021	Sinha et al	BERT, TextRank, TF-IDF, Word2vec, Doc2vev, GloVe, k-means clustering	BBC News Dataset (2225) 5 categories	Comparison of different algorithms for clustering documents based on their content	Dimensionality reduction can lead to information loss. Small dataset is used.	Yes	Yes	-	F1-Score, Precision, Recall
2022	Jouda et al	LexRank, LSA, BART, Pegasus	CNN/DM news Dataset, BART Based data, wiki-how dataset	Comparison of text summarization: abstractive and extractive approaches. Effect of dialects on text summarization techniques. Developed a hybrid system to evaluate the performance of summarization algorithms.	Only sports and politics articles from the CNN/DM dataset were considered. Human evaluators assessed algorithm performance, but their judgments may vary over time due to evolving task understanding or fatigue, introducing variability.	No	Yes	-	ROUGE, BLEU
2022	Kumar et al	Multinomial Naïve Bayes, IDF, LR, RF, SVM,KNN	1800 articles collected from The Hindu, NDTV, Indian Express website.	Proposes an efficient machine learning model for news categorization, introducing a novel IDF integrated MNB model with high accuracy (87.22%)	Relies on traditional machine learning methods. Small Dataset is used.	Yes	No	87.22	Confusion matrix, accuracy, recall, F1-score
2020	Shahin et al	Genism, fastText, Bi-LSTM	News articles collected from different news platforms in bangladesh	Performed Bangla news headline classification	The dataset's focus on Bangla news may constrain the model's applicability to other languages or contexts. Limited exploration of alternative models or methods	Yes	No	85.14	-
2021	Salehin et al	Machine Learning Models, Neural Network models, LSTM	News articles collected from different news platforms in bangladesh	Performed Bangla news classification	The dataset's focus on Bangla news may constrain the model's applicability to other languages or contexts.	Yes	No	-	-

The entire literature review table is present in attached excel sheet

the context of the overwhelming volume of information on online news platforms, ensuring personalized information consumption for readers?

2) RQ2: "What insights can be derived from the exploration of diverse word embedding techniques, such as GloVe, Word2Vec, CBOW, Skip-gram, TF-IDF, and Fasttext, when integrated with machine learning models for news classification? How do these combinations contribute to improving the efficiency of news classification in the digital news landscape?"
3) RQ3: In the realm of extractive summarization techniques, specifically Textrank and LexRank, how can their comparative effectiveness be measured in condensing news articles to provide readers with a succinct yet informative overview of global events, addressing the challenges posed by time constraints and the overwhelming nature of online news?

Our contribution and novelty lie in the comprehensive exploration of these research questions, which collectively bridge the gap between exploring diverse machine learning and word embedding techniques to find the optimal method for the news categorization task, enabling readers to read news according to their preferences. By harnessing extractive summarization techniques like Textrank and LexRank, our study provides

a holistic approach to perform both classification and summarization tasks, thereby contributing to allowing readers to consume news according to their preferences and in less time.

C. Problem Statement

In the era of digital technology, people prefer to access information through online sources rather than reading paper-based text information. These platforms offer a wealth of information on global events, enabling individuals to stay informed from anywhere. While various online news platforms deliver comprehensive content across different fields such as sports and entertainment, not all readers are interested in every category. Many prefer personalized news tailored to their preferences, allowing them to gather more information in less time without delving into lengthy articles. This study addresses this challenge by conducting news classification and summarization using a datasets comprising more than 40,000 news articles spanning diverse fields. The research explores a range of machine learning algorithms and word embedding techniques to identify an optimal algorithmic stack for effective news categorization. Additionally, extractive summarization techniques are examined to provide concise and categorized news articles, delivering information according to individual reader preferences and ensuring the retention of essential details.

D. Novelty of this study

Previous research has primarily concentrated on executing news classification and summarization tasks using machine learning and deep learning approaches; however, most of these endeavors have singularly addressed either the challenge of news classification or news summarization. Unlike Prior works that explored a subset of machine learning and word embedding techniques for performing the news classification task, the uniqueness of this study lies in its comprehensive approach, simultaneously tackling both news categorization and news summarization tasks. Leveraging substantial datasets comprising more than 40,000 news articles that collectively span diverse fields such as sports, business, and entertainment, this research explores a range of machine learning models. These models include Support Vector Machine (SVM), Logistic Regression (LR), Multinomial Naïve Bayes (MNB), Random Forest (RF), and Decision Tree Classifier (DTC), coupled with various word embedding techniques such as GloVe, word2vec, etc. The study aims to discern the optimal algorithmic stack that yields the best results in news classification. Additionally, extractive summarization techniques Textrank and LexRank are scrutinized to to analyze and identify the most effective method for summarizing news articles. This multifaceted approach contributes to bridging the gaps identified in previous research and provides a novel and refined methodology for addressing the dual challenge of news classification and summarization.

E. Significance of Our Work

This study holds paramount significance in the realm of news categorization and summarization, as it systematically explores the intricate interplay between five distinct machine learning classifiers—namely Support Vector Machine (SVM), Logistic Regression (LR), Multinomial Naïve Bayes (MNB), Random Forest (RF), and Decision Tree Classifier (DTC)—and four diverse word embedding techniques—Word2Vec, GloVe, Fasttext, and TF-IDF. The exhaustive experimentation involving the combination of each word embedding technique with every machine learning classifier contributes to a comprehensive understanding of their synergistic effects in the context of news classification. The dataset, amalgamated from CNN news articles spanning from 2011 to 2022 and the News Article Category Test, comprises over 40,000 news articles across 17 distinct categories. This extensive dataset serves as a robust foundation for evaluating the classifiers' performance and elucidating the intricate nuances of news categorization tasks. Furthermore, the application of SVM and LR with TF-IDF as the word embedding technique, yielding superior results, underscores the importance of careful model selection and feature representation in optimizing news classification accuracy. In addition, the incorporation of extractive summarization techniques, specifically LexRank and Textrank on the BBC NEWS summary dataset of over 2000 news articles further enhances the breadth of this research. This work not only contributes valuable insights to the field of news processing but also establishes a benchmark for future endeavors seeking to refine and innovate in the realms of news classification and summarization.

IV. Methodology

This section discusses the various datasets employed for the tasks of news article categorization and summarization using extractive summarization algorithms. The features of the datasets are comprehensively explained, and the methodology is presented in a step-by-step manner for clarity.

A. Dataset

The dataset used in this study for analyzing extractive summarization algorithms, particularly for news summarization, is sourced from the BBC news website.This dataset was initially curated curated for the purpose of data categorization. Comprising a total of 2,225 documents spanning the years 2004-2005, the dataset encompasses news stories distributed across five distinct topical domains, including politics, tech, sport, entertainment, and business. This collection served as the foundational corpus for the research conducted by Greene et al [27] in their study. The dataset not only comprises the articles themselves but also incorporates their corresponding summaries, providing a comprehensive set of textual data for analysis. The inclusion of article summaries makes it particularly suitable for the examination of extractive summarization techniques. Figure 2 visually illustrates the distribution of articles across the five designated categories, while Figure 6 provides valuable insights into the length of articles and summaries within each respective category. This dataset, conveniently accessible on Kaggle, stands as a valuable resource for the thorough investigation and understanding of diverse

news articles spanning multiple fields.

In this study, a comprehensive dataset for news categorization is assembled by combining two distinct sources: the CNN news articles dataset spanning the years 2011 to 2022 and the news article category dataset featuring unique entries from HuffPost. The CNN dataset, comprising approximately 38,000 lines of articles, was meticulously curated using a web crawler. The crawler systematically navigates through the CNN website, extracting a myriad of parameters from the HTML of the articles. This dataset encompasses a rich set of information, including Author, Publication date, Category, Article Section, Url source, Headline, Description, Full text, and more, thereby providing a comprehensive foundation for analysis. On the other hand, the news article category dataset, with 6,877 unique entries, specifically pertains to articles published in HuffPost, categorizing them into diverse topics such as ARTS & CULTURE, BUSINESS, COMEDY, CRIME, EDUCATION, ENTERTAINMENT, ENVIRONMENT, MEDIA, POLITICS, RELIGION, SCIENCE, SPORTS, TECH, and WOMEN. The integration of these datasets aims to leverage the 'article' and 'category' columns for news categorization, employing various machine learning algorithms and word embedding techniques. The study endeavors to determine the optimal combination of these methods, offering valuable insights into effective strategies for news article categorization through a robust and diverse dataset. Figure 6 shows the collective distribution of articles across various categories. The illustration in Figure 5 depicts a representation of each dataset by showcasing a single sample row of data, with emphasis on highlighting the respective column names.

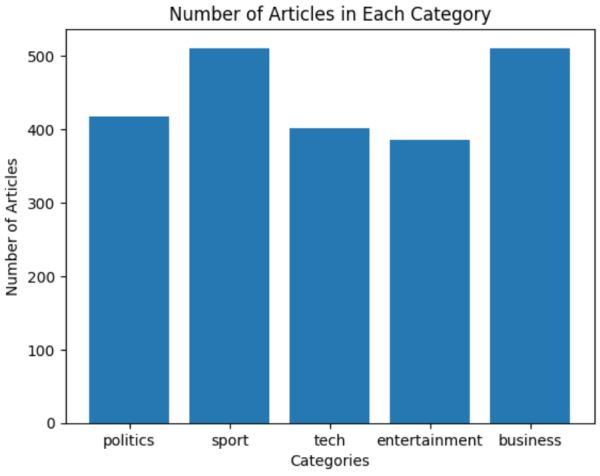

Fig. 2. Number of Articles in Each Category of BBC News Summary Dataset

B. Detailed Methodology

1) Data Preprocessing:: In the initial data preprocessing phase, three datasets—specifically, the CNN News Articles dataset spanning from 2011 to 2022, the News Category Dataset, and the BBC News Summary Dataset are utilized. For the news classification task, the first two datasets are

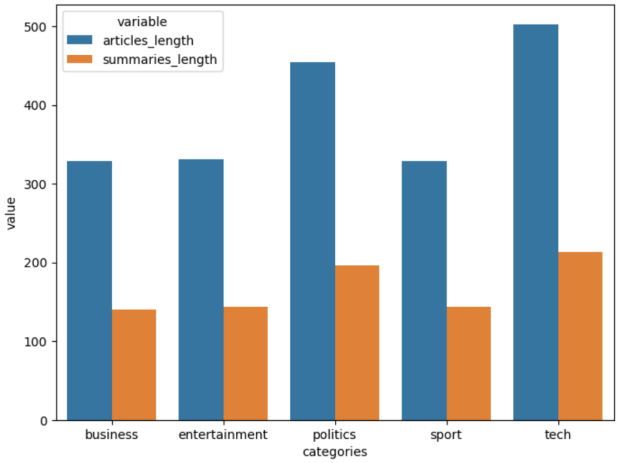

Fig. 3. Length of articles and summaries within each respective category of BBC News Summary Dataset.

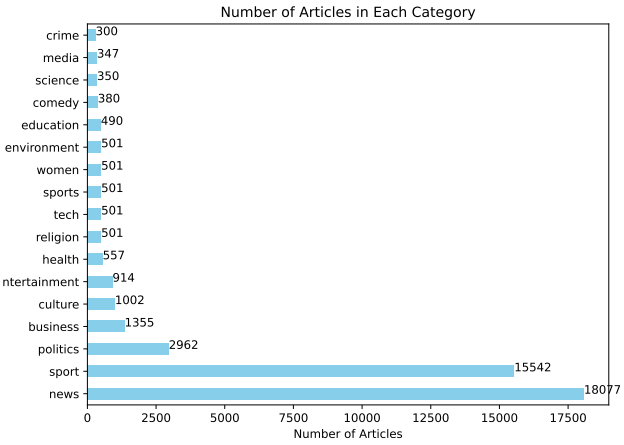

Fig. 4. CNN News Articles from 2011 to 2022 and News Article Category combined dataset's articles distribution across different categories.

integrated, while the third dataset is exclusively set aside for scrutinizing extractive summarization algorithms designed for news summarization. The preprocessing procedures encompass tokenization into words and sentences, followed by the elimination of punctuation, stop words, and other special characters. Subsequent to this, lemmatization is applied to standardize word forms. Following lemmatization, word vectors are generated, and features are extracted using various word embedding techniques such as TF-IDF, GloVe, Fasttext, and Word2Vec. These embeddings serve as input for machine learning classifiers. All these preprocessing tasks are applied to the merged dataset, while the BBC News Summary dataset only undergoes tokenization. Regarding the exploration of extractive summarization techniques for news summarization, graph-based methods like LexRank and Textrank are employed. These techniques necessitate only tokenization as a preprocessing step to transform raw text into standardized

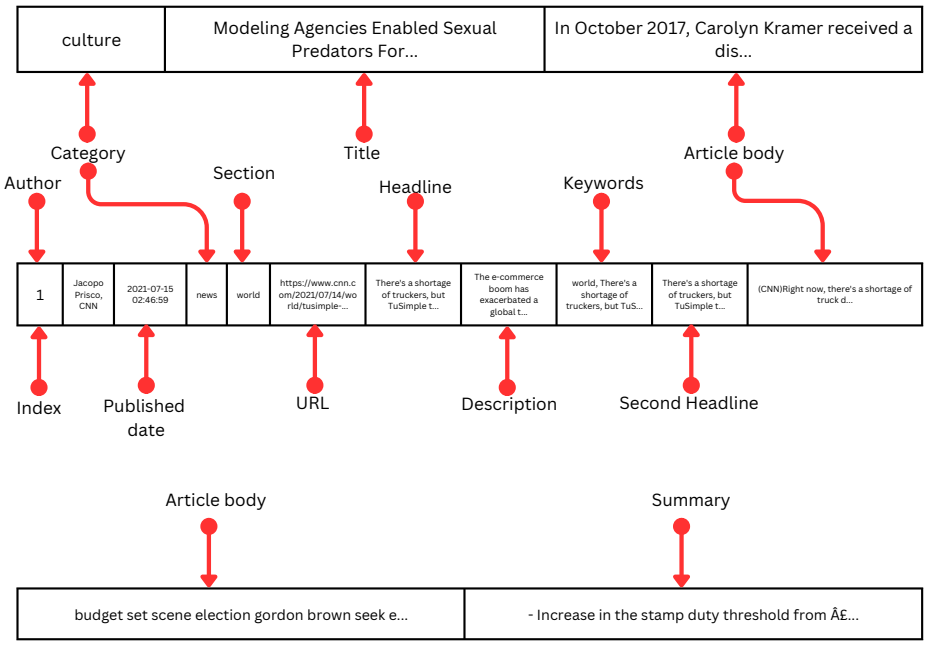

Fig. 5. Sample Datasets Image: The first dataset comprises entries from the News Article Category dataset. The second dataset consists of CNN News Articles spanning the years 2011 to 2022. The third dataset showcases the BBC News Summary dataset.

text suitable for summarization, thereby extracting crucial sentences and summarizing the text.

2) Word Embeddings:: The word embedding phase involves creating vector representations for words through the utilization of TF-IDF, GloVe, FastText, and Word2Vec techniques. Each technique is applied independently, generating distinct sets of word vectors. These embeddings capture the semantic associations between words, enabling the extraction of meaningful features for subsequent classification. TF-IDF assesses the word's significance in the document relative to other documents. Word2Vec, a neural network-based technique, learns distributed word representations. GloVe, akin to Word2Vec, is a method for word representation that also captures semantic relationships. FastText, an extension of Word2Vec, introduces subword information into embeddings. The incorporation of multiple word embedding techniques is intended to explore and leverage diverse linguistic representations inherent in the datasets. The resultant embeddings serve as input features for machine learning classification models.

3) Machine Learning Classification Models: After obtaining word embeddings, the study employs machine learning classifiers, including Support Vector Machine (SVM), Logistic Regression (LR), Multinomial Naive Bayes (MNB), Random Forest (RF), and Decision Tree Classifier (DTC). Each word embedding technique is paired with every classification model to identify the optimal combination for news classification. The data is divided into 80% training and 20% testing sets for model evaluation. Precision, recall, and F1-score are employed as evaluation metrics to assess the performance of each combination, providing insights into the effectiveness of different algorithms for news classification.

4) Extractive Summarization:: In the context of extractive text summarization, the preprocessing of the BBC News Summary Dataset involves tokenization, a crucial step that transforms the raw text of news articles into a format suitable for input into graph-based extractive summarization algorithms. Specifically, the study focuses on analyzing the performance of unsupervised techniques, specifically the Textrank and LexRank algorithms using the BBC News Summary dataset. The initial step in this process is the creation of vector representations, known as word embeddings, for each sentence in the dataset. Subsequently, the study calculates similarities between these sentence vectors, storing the results in a matrix. This matrix is then transformed into a graph, with sentences as vertices and similarity scores as edges. The rank of each sentence is determined within this graph, and the final summary comprises a specific number of top-ranked sentences. To streamline this summarization process, the study leverages the Sumy library, which implements TextRank and LexRank algorithms. Notably, TextRank and LexRank are graph-based extractive summarization algorithms that identify key sentences based on the relationships between sentences in the dataset. ROUGE-1, ROUFE-2, ROUGE-L and BLEU are employed as evaluation metrics to quantitatively assess the quality and coherence of the generated summaries. A figure depicting the workflow of this study (shown in Fig 4

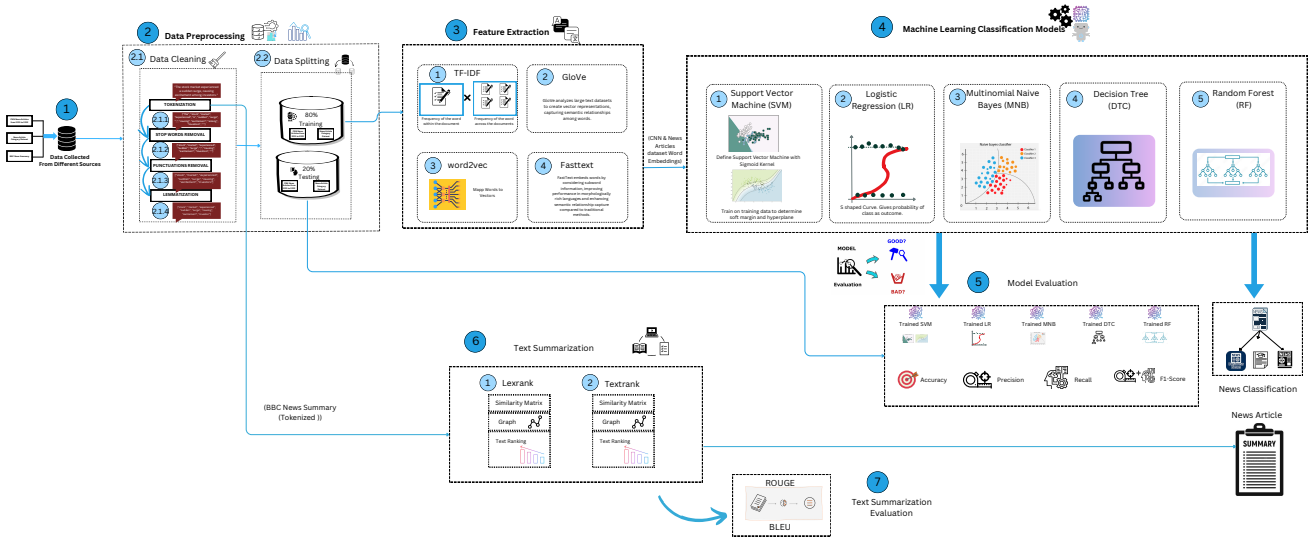

Fig. 6. Workflow. (1) Three datasets are utilized in this study. (2.1) The datasets undergo various cleaning processes, including tokenization, stop word removal, punctuation removal, and lemmatization techniques as part of the data preprocessing section. (2.2) After preprocessing, the combined News Article Category and CNN News Article datasets are divided into training and testing data. (3) The training data is then vectorized using different word embedding techniques. (4) The features extracted by applying different word embedding techniques are utilized by various Machine Learning Classification Models to perform the News Classification Task. (5) Different classification models are evaluated using test data, employing accuracy, precision, recall, and F1-score as evaluation metrics. (6) The BBC News Summary Dataset, after preprocessing, is employed for the news summarization task using Extractive Summarization Techniques. (7) ROUGE and BLEU are employed as evaluation metrics to assess the performance of the news summarization task.

C. Evaluation Metrics

In the course of this investigation, we employ a suite of widely recognized evaluation metrics to systematically appraise the effectiveness of machine learning classifiers and extractive summarization algorithms. These metrics assume a crucial role in objectively measuring the model's competence in managing multifaceted tasks such as multi-class classification and text summarization, providing a comprehensive evaluation through key indicators like precision, recall, and F1-Score.

Below are the evaluation metrices used in this study for evaluation different machine learning classifiers.

1) **Precision:** Precision is the ratio of correctly predicted positive observations to the total predicted positives. In the context of news classification, precision measures the accuracy of the classifier in identifying relevant news articles.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

2) **Recall:** Recall (Sensitivity or True Positive Rate) is the ratio of correctly predicted positive observations to the all observations in actual class. In the news classification scenario, recall signifies the ability of the classifier to capture all relevant news articles.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

3) **F1-Score:** F1-Score is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives. It is particularly useful when there is an uneven class distribution.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

For the evaluation of extractive summarizationa algorithms ROUGE and BLEU are used in this study.

1) **ROUGE-1, ROUGE-2, ROUGE-L:** ROUGE measures the overlap of n-grams (words or sequences of words) between the generated summary and the reference summary. ROUGE-1 considers unigrams, ROUGE-2 considers bigrams, and ROUGE-L measures the longest common subsequence.

$$\text{ROUGE-1} = \frac{\text{Number of Overlapping Unigrams}}{\text{Total Unigrams in Reference Summary}}$$

$$\text{ROUGE-2} = \frac{\text{Number of Overlapping Bigrams}}{\text{Total Bigrams in Reference Summary}}$$

$$\text{ROUGE-L} = \frac{\text{Length of Longest Common Subsequence}}{\text{Total Unigrams in Reference Summary}}$$

2) **BLEU (Bilingual Evaluation Understudy):** BLEU evaluates the quality of machine-generated text by comparing it to one or more reference texts. It considers n-grams precision, where n is usually set to 1, 2, 3, or 4.

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} \frac{1}{N} \cdot \log(p_n)\right)$$

Where BP is the brevity penalty, and p_n is the n-grams precision.

D. Experimental settings

In the experimental setup of our research, we utilized different types of models, including Support Vector Machine (SVM) with a sigmoid kernel, Random Forest, Decision Tree, Logistic Regression, and Multinomial Naive Bayes, for performing a news classification task. To capture non-linear relationships in the information, a sigmoid kernel was applied to SVM. Random Forest was configured with 100 trees, and hyperparameters were fine-tuned for optimal performance. Decision Tree was kept at default settings. Logistic Regression used the 'lbfgs' solver and underwent a maximum of a thousand iterations. Multinomial Naive Bayes was also included in our comparative analysis. The experiments were conducted with a balanced class-weight setting to address potential class imbalances. To ensure reproducibility, random seed values were fixed, and significant performance metrics such as accuracy, precision, recall, and F1-score were employed for model evaluation. Our experimental framework aimed to provide a comprehensive assessment of the proposed models against each other and competing methods, ensuring a robust comparison of their classification performance.

TABLE II
BLEU AND ROUGE SCORES FOR TEXTRANK AND LEXRANK

Algorithm	BLEU Score	ROUGE-1	ROUGE-2	ROUGE-L
Textrank	0.28	0.52	0.41	0.51
LexRank	0.20	0.52	0.42	0.52

V. RESULTS

1) RQ1: "Optimizing News Classification for Personalized Information Consumption: Leveraging Diverse ML Models (SVM, LR, MNB, RF, DTC)"
The study's results underscore the effectiveness of machine learning classifiers, specifically SVM, LR, MNB, RF, and DTC, in the task of news article categorization. Addressing the first research question – "How can a diverse range of machine learning models, including SVM, LR, MNB, RF, and DTC, be effectively employed to optimize news classification in the context of the overwhelming volume of information on online news platforms, ensuring personalized information consumption for readers?" Figures 7, 8, 9, 10 present the outcomes of various algorithmic stacks used for the news categorization task. Through analysis, it is evident that Support Vector Machine and Logistic Regression machine learning classifiers deliver remarkable results in news classification. This outcome supports the objective of efficiently managing unstructured documents, enabling readers to tailor their news consumption based on individual preferences.
2) RQ2:"Enhancing News Classification Efficiency: Integrating Word Embedding Techniques with ML Models" In addressing the second research question, this study utilized different word embedding techniques, including Word2Vec, GloVe, FastText, and TF-IDF, in conjunction with various machine learning classifiers such as Support Vector Machine (SVM), Logistic Regression (LR), Support Vector (SV), Multinomial Naive Bayes (MNB), Random Forest (RF), and Decision Tree Classifier (DTC). We explored every combination of word embedding technique with each machine learning classifier. Figures 7, 8, 9, 10 show the results of different algorithmic stacks used for the news categorization task. Through the examination of these figures and evaluation metrics such as Precision and Recall, it becomes evident that the combination of the Term Frequency-Inverse Document Frequency (TF-IDF) word embedding technique with different machine learning classifiers yields remarkable results compared to other word embedding techniques. By experimenting with various word embedding techniques, we sought to identify the optimal algorithmic stack for the news classification task, ultimately enhancing efficiency in the digital news landscape.
3) RQ3:"Assessing Textrank and LexRank: Comparative Effectiveness in Condensing News Articles for Concise and Informative Overviews amid Time Constraints and Online News Overwhelm" To comprehensively address the third research question, this study utilizes the BBC News Summary dataset to analyze the performance of two graph-based extractive summarization techniques, Textrank and LexRank, using ROUGE-1, ROUGE-2, ROUGE-L, and BLEU evaluation metrics. Table II displays the values of these evaluation metrics for both the Textrank and LexRank algorithms. Textrank has a higher BLEU score, while the ROUGE scores for both algorithms are the same. Considering the combined information from BLEU and ROUGE, Textrank appears to perform slightly better in terms of n-gram overlap and summarization quality. Thus, by using these two extractive summarization algorithms, we can generate extractive summaries of news articles and overcome the issues of time constraints and news overwhelm.

VI. DISCUSSION

In the pursuit of comprehensively addressing the tasks of news categorization and summarization, our research strategically employed five powerful machine learning classi-

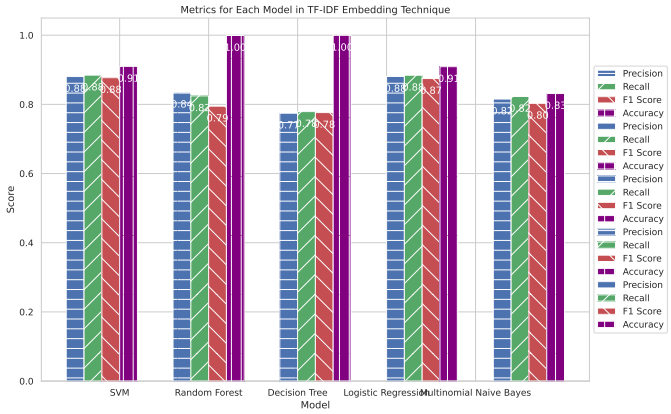

Fig. 7. Comparison of key metrics: precision, recall, F1 Score, and training accuracy across various models using TF-IDF embeddings.

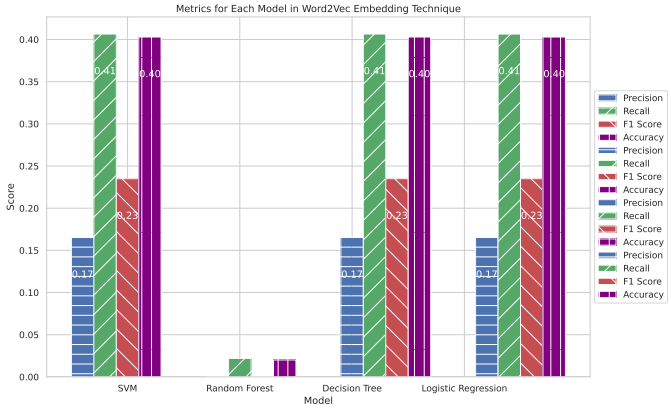

Fig. 8. Comparison of key metrics: precision, recall, F1 Score, and training accuracy across various models using word2vec embeddings.

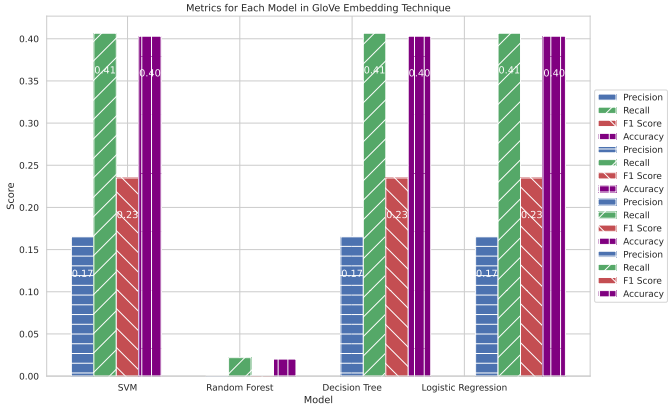

Fig. 9. Comparison of key metrics: precision, recall, F1 Score, and training accuracy across various models using GloVe embeddings.

fiers—Support Vector Machine (SVM), Random Forest (RF), Decision Tree Classifier (DTC), Logistic Regression (LR), and Multinomial Naive Bayes (MNB). The primary goal was to optimize news classification for personalized information

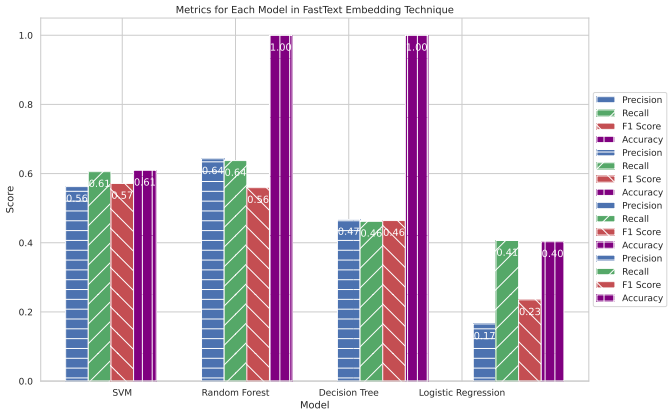

Fig. 10. Comparison of key metrics: precision, recall, F1 Score, and training accuracy across various models using Fasttext embeddings.

consumption, reflecting the evolving nature of digital media consumption. To achieve this, our approach embraced five distinct word embedding techniques, providing the vectors generated by each technique to every classifier. This systematic methodology aimed to discern the most effective word embedding-machine classifier combinations, ensuring a nuanced understanding of how different models interact with diverse embedding strategies. The amalgamation of two extensive datasets, encompassing CNN News Articles from 2011 to 2022 and the News Article Category dataset, resulted in a substantial corpus exceeding 40,000 articles spanning seventeen distinct categories, thereby enriching the diversity and depth of our analysis. For the critical task of news summarization, our research leveraged the BBACNews Summary dataset, subjecting it to a meticulous evaluation of two prominent graph-based unsupervised extractive summarization algorithms: Textrank and LexRank. With over 2000 news articles representing various fields, this dataset offered an ideal foundation for an in-depth analysis of extractive summarization algorithms, ensuring the robustness and generalizability of our findings.

Turning to our first research question, "Optimizing News Classification for Personalized Information Consumption," the outcomes of various algorithmic stacks for news categorization were meticulously presented in Figures 7, 8, 9, 10. Notably, SVM and LR emerged as standout performers, showcasing remarkable precision and recall values of 0.88%. These results underscored their effectiveness in addressing the intricate challenge of news categorization, positioning our study as pivotal in empowering readers to tailor their news consumption according to their preferences, thereby enhancing the user experience.

The second research question, "Enhancing News Classification Efficiency: Integrating Word Embedding Techniques with ML Models," delved into the nuanced realm of different word embedding techniques—word2vec, TF-IDF, GloVe, and Fasttext. Our thorough analysis of machine learning classifiers' performance, considering each embedding technique, unveiled

that classifiers utilizing TF-IDF as input exhibited superior and remarkable results compared to other embedding techniques. This finding emphasizes the critical role of embedding strategies in optimizing the efficiency of news classification models, offering valuable insights for practitioners and researchers alike.

Addressing the third research question, "Assessing Textrank and LexRank: Comparative Effectiveness in Condensing News Articles," our research conducted a comprehensive comparison of Textrank and LexRank using various evaluation metrics. Textrank consistently outperformed LexRank across all metrics, delivering concise and informative overviews within the constraints of time and the overwhelming nature of online news. These findings highlight the practical significance of Textrank in the context of extractive summarization, suggesting its preference for effectively condensing news articles for readers with time constraints.

Our investigation further revealed that MNB yielded results exclusively using TF-IDF, as other embedding techniques produced negative vectors unsuitable for MNB input. This underscores the necessity for continued exploration and refinement of word embedding techniques to cater to the unique requirements of diverse classifiers.

In conclusion, our research significantly contributes to the fields of news categorization and summarization, providing a robust foundation for future studies. The holistic approach, encompassing multiple classifiers, embedding techniques, and extractive summarization algorithms, offers readers personalized and succinct news aligned with their preferences. These findings not only advance the current understanding of news processing but also highlight avenues for continued exploration and refinement in the quest for efficient and personalized information consumption in the dynamic landscape of digital news. The evolving nature of media consumption demands ongoing research to stay at the forefront of technological advancements and user preferences, ensuring the continued relevance and impact of studies in this domain.

A. Limitations

Limitations are inherent in this study as it primarily concentrates on the exploration and application of diverse combinations of machine learning classifiers, including Support Vector Machine (SVM), Logistic Regression (LR), Multinomial Naïve Bayes, Random Forest, and Decision Tree Classifier, coupled with various word embedding techniques such as TF-IDF, WORD2VEC, GloVe, and Fasttext. However, it is crucial to note that deep learning techniques have been excluded from our discussion and implementation. In the realm of text summarization, our investigation is confined to extractive techniques, specifically LexRank and Textrank. The study does not extend its purview to encompass abstractive text summarization methods. It is essential to acknowledge that this research does not delve into the comprehensive exploration or discussion of abstractive summarization techniques. The limitations of the current work underscore the need for future research endeavors to encompass a broader spectrum of machine learning classifiers, deep learning methodologies, and both extractive and abstractive summarization techniques for a more comprehensive understanding of text processing and summarization tasks.

B. Future Directions

The outcomes of our research present promising trajectories for future investigations within the realm of news categorization and summarization. There exists a significant opportunity for delving into the realm of deep learning techniques to enhance feature extraction and classification processes for news articles, complementing the array of machine learning classifiers explored in our study. A prospective avenue for further exploration lies in the investigation of abstractive summarization algorithms to advance the automated summarization of news articles, offering a nuanced approach beyond the extractive techniques examined in the current study. Additionally, the integration of sentiment analysis into news articles presents an intriguing area for future research. By tailoring news content to readers' preferences, this approach aims to streamline information dissemination, optimize time utilization, and present readers with condensed yet informative sentences. This personalized approach is envisioned to empower readers with a more efficient and tailored means of acquiring substantial information within a condensed timeframe. As we contemplate the future trajectory of this study, it is evident that these directions hold the potential to enrich and expand the understanding of news processing and enhance the delivery of relevant information to diverse readership.

VII. Conclusion

In conclusion, the comprehensive experimentation involving various machine learning algorithms and word embedding techniques has yielded valuable insights into the domain of news categorization and summarization. Among the diverse algorithms considered, TF-IDF combined with SVM and LR emerged as a formidable stack, demonstrating exceptional performance in the classification of news articles across different categories. This finding highlights the efficacy of employing traditional text representation methods in conjunction with well-established classifiers for improved accuracy and relevance. Additionally, the incorporation of extractive summarization techniques, specifically TextRank, has proven to be highly effective in generating concise and informative summaries for news articles. The success of TextRank in comparison to LexRank underscores its superiority in the unsupervised graph-based summarization domain. Overall, these results contribute significantly to the advancement of news-oriented natural language processing applications, offering readers the ability to access tailored news content efficiently and enhancing their overall reading experience.

References

[1] H. UNGER, "Cnn news articles from 2011 to 2022," https://www.kaggle.com/datasets/hadasu92/cnn-articles-after-basic-cleaning, 2021.

[2] B. Timilsina, "News article category dataset," 08 2021.

[3] P. SHARIF, "Bbc news summary," https://www.kaggle.com/datasets/pariza/bbc-news-summary, 2017.

[4] S. Kanaujia and R. Srivastava, "Online newspapers: News on fast track," *Library Hi Tech News*, vol. 20, no. 10, 2003.

[5] E. Ortiz-Ospina and M. Roser. (2023) The rise of social media. Our World in Data. [Online]. Available: https://ourworldindata.org/rise-of-social-media

[6] C. S. Park and B. K. Kaye, *What's this? Incidental Exposure to News on Social Media, News-Finds-Me Perception, News Efficacy, and News Consumption*. Routledge, 2021, pp. 98–121.

[7] K. I. Malik, "Urdu news content classification using machine learning algorithms," *Lahore Garrison University Research Journal of Computer Science and Information Technology*, vol. 6, no. 1, pp. 22–31, 2022.

[8] M. Kayakuş and F. Y. AÇIKG¨OZ, "Classification of news texts by categories using machine learning methods," *Alphanumeric Journal*, vol. 10, no. 2, pp. 155–166, 2022.

[9] B. Juarto *et al.*, "Indonesian news classification using indobert," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 2, pp. 454–460, 2023.

[10] R. Wang, Z. Li, J. Cao, T. Chen, and L. Wang, "Convolutional recurrent neural networks for text classification," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–6.

[11] A. Jouda, "Exploring abstractive vs. extractive summarisation techniques for sports news," in *Irish Conference on Artificial Intelligence and Cognitive Science*. Springer, 2022, pp. 368–380.

[12] S. Deo and D. Banik, "Text summarization using textrank and lexrank through latent semantic analysis," in *2022 OITS International Conference on Information Technology (OCIT)*. IEEE, 2022, pp. 113–118.

[13] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, "Neural abstractive text summarization with sequence-to-sequence models," *ACM Transactions on Data Science*, vol. 2, no. 1, pp. 1–37, 2021.

[14] L. Sinha and S. J. Nirmala, "Combining document embedding techniques for clustering and analysis of extractive summaries," in *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. IEEE, 2021, pp. 1–7.

[15] D. B. Bisandu, R. Prasad, and M. M. Liman, "Clustering news articles using efficient similarity measure and n-grams," *International Journal of Knowledge Engineering and Data Mining*, vol. 5, no. 4, pp. 333–348, 2018.

[16] B. Jang, I. Kim, and J. W. Kim, "Word2vec convolutional neural networks for classification of news articles and tweets," *PloS one*, vol. 14, no. 8, p. e0220976, 2019.

[17] S. Kumar, A. Sharma, B. K. Reddy, S. Sachan, V. Jain, and J. Singh, "An intelligent model based on integrated inverse document frequency and multinomial naive bayes for current affairs news categorisation," *International journal of system assurance engineering and management*, vol. 13, no. 3, pp. 1341–1355, 2022.

[18] M. M. H. Shahin, T. Ahmmed, S. H. Piyal, and M. Shopon, "Classification of bangla news articles using bidirectional long short term memory," in *2020 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 2020, pp. 1547–1551.

[19] K. Salehin, M. K. Alam, M. A. Nabi, F. Ahmed, and F. B. Ashraf, "A comparative study of different text classification approaches for bangla news classification," in *2021 24th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2021, pp. 1–6.

[20] I. Rabbimov and S. Kobilov, "Multi-class text classification of uzbek news articles using machine learning," in *Journal of Physics: Conference Series*, vol. 1546, no. 1. IOP Publishing, 2020, p. 012097.

[21] X. Chen, P. Cong, and S. Lv, "A long-text classification method of chinese news based on bert and cnn," *IEEE Access*, vol. 10, pp. 34 046–34 057, 2022.

[22] I. K. Izzah and A. S. Girsang, "Modified tf-assoc term weighting method for text classification on news dataset from twitter," *IAENG International Journal of Computer Science*, vol. 48, no. 1, pp. 142–151, 2021.

[23] M. Tomer and M. Kumar, "Stv-beats: Skip thought vector and bi-encoder based automatic text summarizer," *Knowledge-Based Systems*, vol. 240, p. 108108, 2022.

[24] Y. Du and H. Huo, "News text summarization based on multi-feature and fuzzy logic," *IEEE Access*, vol. 8, pp. 140 261–140 272, 2020.

[25] A. K. Singh and M. Shashi, "Vectorization of text documents for identifying unifiable news articles," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 7, 2019.

[26] Y. Chen and Q. Song, "News text summarization method based on bart-textrank model," in *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. IEEE, 2021, pp. 2005–2010.

[27] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 377–384.