iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XI **Month of publication:** November 2024

DOI: https://doi.org/10.22214/ijraset.2024.65443

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XI Nov 2024- Available at www.ijraset.com

An Automated Daily News Reports Generating Application Involving Keyword-Based News Scraping, Summarization, and Sentimental Analysis Leveraging NLP Models

Nafeesa Begum J[1], Chandru D[2]
[1]*Professor*, [2]*Final Year Student, Department of Computer Science and Engineering, Government College of Engineering, Bargur, Krishnagiri, Tamilnadu, India*

Abstract*: *In the era of information overload, obtaining relevant, concise, and sentiment-analyzed news content is essential for effective decision-making. This paper introduces an automated daily news reporting system that streamlines the process of collecting, summarizing, and analyzing the sentiment of news articles fetched based on the user's keyword. By leveraging state-of-the-art Natural Language Processing (NLP) models like GPT-3.5 for two-level summarization and BERTweet for sentiment analysis, the system provides users with concise, sentiment-labeled reports, enhancing their understanding of news trends and emotional tone.

The architecture integrates data scraping, text extraction, and sentiment classification within a cloud-based Python microservice, supported by Flask. The system incorporates user localization options, allowing users to customize news retrieval by region and time preferences.

Finalized reports are formatted into HTML and delivered directly to authenticated user emails through Gmail API, ensuring seamless and secure distribution. This research underscores the significance of automated news summarization and sentiment analysis in modern information retrieval. It provides a scalable and personalized solution that enables real-time synthesis of large volumes of news content, making it accessible and relevant for diverse audiences worldwide.

Keywords*: *News Summarization, GPT-3.5, BERTweet, Transfer Learning, Sentiment Analysis, NLP, Keyword-based Scraping

I. INTRODUCTION

In today's fast-paced world, staying informed about the latest developments is critical, yet the overwhelming amount of information can make it difficult for users to focus on relevant news. With the rapid growth of online content, automated systems for news summarization and sentiment analysis have become essential tools for enhancing information accessibility and supporting decision-making.

This paper introduces a novel automated daily news report generator that combines keyword-based news scraping, content summarization, and sentiment analysis to deliver concise, sentiment-labeled news summaries directly to users' emails. The project employs advanced Natural Language Processing (NLP) models to handle large volumes of news data efficiently. Specifically, we utilize the GPT-3.5-turbo-0125 model for summarization and the BERTweet model for sentiment analysis. Both models leverage transfer learning, a method that allows the application of pre-trained models on large datasets, fine-tuned for specific tasks to maximize performance while minimizing computational overhead. This approach enhances adaptability and reduces the need for extensive retraining, enabling the system to effectively process diverse news content. The core objectives of this research are to develop a robust, automated pipeline that can retrieve, summarize, and analyze news articles based on user-defined keywords. Each news article is labeled with sentiment indicators (positive, negative, or neutral), providing a personalized reading experience. The project integrates a cloud-based Python microservice for data scraping, summarization, and sentiment analysis, along with an interface for user localization preferences. Reports are delivered directly to users' authenticated emails through the Gmail API, providing a seamless and secure distribution process. This system not only enhances the accessibility of real-time news but also enables users to receive news summaries in line with their interests and emotional preferences, illustrating the transformative role of NLP in modern information retrieval.

II. LITERATURE SURVEY

CoVShorts is a web application that uses transformer models (BERT, GPT-2, XLNet, BART, T5) to summarize COVID-19 news articles. BERT outperforms others in generating the most relevant summaries based on ROUGE scores. The tool helps users access concise and accurate summaries, validated by Word Cloud analysis. Future enhancements include a mobile app and URL-based input for automatic summarization [1]. The increasing volume of text data has led to a rising demand for automatic text summarization, which can be divided into extractive and abstractive methods. Extractive summarization is accurate but lacks coherence, while abstractive methods generate readable summaries but may omit important information. Recent advances in deep learning, particularly in multi-document summarization tasks, have improved the quality of abstractive summarization. Combining both approaches in hybrid models has shown promising results for better information retention and readability [2]. A comprehensive review of Automatic Text Summarization (ATS) methods is provided, with a focus on techniques based on Large Language Models (LLMs). The paper outlines key stages of the ATS process, such as data acquisition, pre-processing, and evaluation metrics, and compares traditional extractive and abstractive methods with newer LLM-based approaches. LLMs are shown to enhance summary quality and flexibility. Future research directions involve improving prompt design, addressing domain-specific challenges, and optimizing LLM-generated outputs [3].

An AI-based tool using GPT-3.5 Turbo 16k and Pegasus for automatic summarization of news articles has been developed. The tool employs web scraping, article similarity calculations, and audio matching to produce concise summaries. Evaluation with BLEU and ROUGE metrics indicates that GPT-3.5 Turbo outperforms the Pegasus model. Future improvements aim to add multilingual support and incorporate audio transcriptions, increasing the accessibility and relevance of online news content [4]. GPT-3.5, GPT-4, and Llama 2 were compared in sentiment analysis tasks, and the results revealed that LLMs, even in zero-shot settings, can match or surpass traditional transfer learning models in sentiment classification accuracy. Among the models, Llama 2 provided the most explainable sentiment classifications. This study suggests that LLMs are particularly useful for marketing research, although challenges such as bias and reproducibility remain [5].

A detailed overview of Automatic Text Summarization (ATS) techniques is provided, including extractive, abstractive, and hybrid approaches. The paper discusses the shift from rule-based methods to deep learning models like GPT, highlighting the challenges in improving summary accuracy, handling domain-specific terminology, and maintaining coherence. The review also addresses the need for real-time summarization and ethical concerns such as bias mitigation. Future research should focus on optimizing ATS for large datasets and real-time applications [6]. A web application has been developed to improve academic literature searches by automating the process using web scraping and crawling techniques. This tool reduces the time researchers spend manually reviewing literature and enhances accessibility to relevant studies. Key features include broader coverage, faster retrieval of abstracts, and integration of legal and ethical considerations. Future improvements will include AI-driven suggestions and advanced data analytics to optimize the search process [7].

A comprehensive review of sentiment analysis is presented, focusing on recent advancements in machine learning, deep learning, and large language models (LLMs). The paper explores various application domains, common datasets, and evaluation metrics. It highlights the strengths and limitations of current approaches and discusses the challenges faced by existing models. The review suggests future research directions to overcome these challenges and improve sentiment analysis techniques, emphasizing their importance in data-driven decision-making [8]. A news article summarization system has been proposed, which automatically gathers and summarizes content from local online newspapers using a custom-built web crawler. The system employs computational linguistic techniques such as triplet extraction, semantic similarity, and OPTICS clustering with DBSCAN to generate coherent summaries. The performance is evaluated using the ROUGE metric, showing strong results. This approach addresses news overload by condensing large volumes of information into concise, readable summaries [9].

The paper [10] investigates the use of Pre-trained Models (PTMs) like BERT, GPT, and T5 for various NLP tasks, including sentiment analysis, news classification, and anti-spam detection. PTMs offer significant benefits, including data efficiency and generalization across domains, though challenges remain in terms of computational cost, overfitting, and interpretability. The paper emphasizes fine-tuning, transfer learning, and anomaly detection techniques. Despite these challenges, PTMs are becoming increasingly important in modern NLP applications. Sentiment analysis applied to news articles is explored, with a focus on the challenges of detecting sentiment due to variability in sentiment words and the complexity of negations. The paper compares lexicon-based methods with machine learning techniques for sentiment detection. It highlights the limitations of current approaches in handling complex sentence structures and language-specific issues. Future work should aim to expand sentiment analysis to other languages and refine feature handling techniques [11].

Pre-trained transformer models like BERT, XLNet, and DistilBERT were evaluated for sentiment analysis on the IMDb reviews dataset. XLNet performed better than the others, particularly in handling token dependencies. DistilBERT offered faster training times while still delivering strong performance [12]. The paper [13] introduces an enhanced BERT-based Convolution Bi-directional Recurrent Neural Network (CBRNN) model for sentence-level sentiment analysis, addressing challenges such as noisy data and contextual loss. The model uses zero-shot classification for annotating data and applies BERT for semantic and contextual feature extraction. It outperforms traditional word embedding models like GloVe and Word2Vec on multiple datasets, improving accuracy, F1-score, and AUC. Future work will focus on applying the model to resource-poor languages and multi-class classification. An Automated news article summarization system based on keyword queries was developed and compared using the BART and T5 transformer models. BART outperformed T5 in terms of F1 score, precision, and recall. The system uses web scraping to filter out irrelevant content, allowing for efficient summarization of mid-sized news articles. Future improvements include refining the model for more complex articles and expanding its capabilities [14]. AI-driven NLP models like BERT and GPT were explored for multilingual text summarization and translation. The study highlighted the importance of hyperparameter optimization, with the best performance achieved using a learning rate of 3×10^{-5}. BERT achieved a ROUGE-1 score of 0.50 for summarization, while GPT attained a BLEU score of 38.7 for translation. The research suggests that further work on low-resource languages is necessary and emphasizes AI's potential in transforming sectors like healthcare and law [15].

III. METHODOLOGY

A. Application Architecture

Fig. 1 presents the architectural framework of the mobile application, designed for efficient data processing, user accessibility, and automated report generation. The application is built using the React Native framework, chosen for its cross-platform compatibility and responsive user interface. State management within the application is efficiently handled through the Redux library, ensuring that component states remain consistent and optimized across user sessions. Data management is facilitated by storing collected information in a Pandas DataFrame, which provides a structured and efficient means of handling large volumes of data. This data frame serves as the central repository, enabling easy manipulation and seamless integration throughout the data processing pipeline. For user authentication, the application employs Firebase Authentication, allowing users to securely log in using their Google accounts. This integration provides both convenience and security, ensuring that only authenticated users can access personalized news summaries.

The core processing pipeline is responsible for automating data scraping, content extraction, summarization, sentiment analysis, and email distribution is implemented as a cloud-hosted Python microservice built with the Flask web framework. This backend service is powered by various specialized libraries and models, including GoogleNews and Newspaper4k for data scraping and extraction, GPT-3.5-turbo for text summarization, and the BERTweet model for sentiment analysis. The pipeline also includes Gmail API configuration to facilitate seamless email delivery. The cloud-hosted nature of this microservice ensures high accessibility, allowing users to retrieve news summaries and sentiment-labeled content in real time, irrespective of their location. This scalable architecture is designed to handle high volumes of news data, making it suitable for users seeking efficient, personalized news delivery.

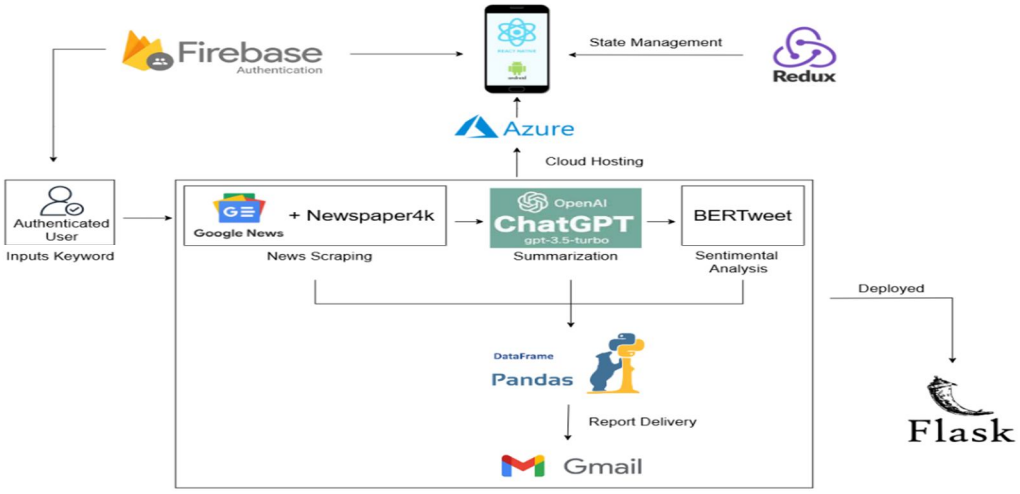


Fig. 1 The application design architecture

B. Interface Localization

The application's interface includes user authentication through Google accounts, ensuring a secure and personalized experience. Once authenticated, users can configure their profile settings to select specific regions (e.g., India, United States) and a preferred time period for news collection. These preferences, which are stored by default in the user's profile, allow for flexible and personalized news filtering based on location and recency. Users can modify their regional and time-based preferences at any time to adjust the scope and relevance of the news content retrieved.

Upon entering a keyword, the application scrapes and filters news articles based on these preferences, ensuring that only relevant articles from the selected region and time period are included. The final report, containing a concise summary of each article, along with details such as Title, Media Source, Timestamp, Description, Link, and Sentiment Label, is then delivered directly to the user's authenticated email. This functionality not only provides timely news tailored to user interests but also enhances the utility of the application by offering a targeted, sentiment-labeled overview of news content.

C. Pipeline processing

The pipeline process is integrated into the architectural design as a cloud-based microservice developed using Python's Flask web framework. This microservice acts as the backend for automating the entire workflow which supports a full cycle of news collection and extraction, summarization, sentiment analysis, and distribution of reports capable of handling end-to-end processing for automated news analysis, making it a powerful tool for real-time information retrieval and to create structured reports as outlined below.

1) Data Collection and Extraction

The data collection and extraction process begins with a cloud-based microservice that uses the GoogleNews and Newspaper4k libraries for comprehensive news scraping and content retrieval. GoogleNews enables keyword-based searches [7] based on user-defined criteria such as region and time period, efficiently gathering URLs for news articles related to specific topics. This approach allows users to specify keywords and retrieve a diverse set of articles that align with their preferences, ensuring a rich dataset of relevant news content. Once URLs are collected, Newspaper4k processes each link to scrape the full text, metadata, and other relevant details, while performing essential cleaning tasks like removing HTML tags and handling special characters. This ensures that the data is prepared for further NLP tasks, such as summarization and sentiment analysis, by providing well-structured and clean content. After extraction, the raw news content is stored in a Pandas DataFrame, a structured format ideal for efficient management and future processing. Each row in the DataFrame represents a news article, with columns for the title, media source, timestamp, description, and link, creating a reliable storage system for subsequent steps in the pipeline. This setup not only facilitates efficient data management but also allows for custom configurations, enabling targeted news collection based on user preferences for specific regions and time periods. The combined use of GoogleNews and Newspaper4k libraries supports an automated, flexible, and scalable data collection process, forming the foundation for downstream NLP tasks in the application.

2) News content summarization

Unlike traditional news applications that provide full-length articles, we implement the automated process of summarizing news articles into concise summaries. This saves users time and allows them to quickly grasp the main points of multiple articles without having to read through each one in its entirety. The extracted data is passed to the summarization module that uses GPT-3.5 Turbo [6] for initial and refined summaries, ensuring comprehensive content aggregation. This step is crucial for distilling large amounts of information into concise summaries.

- Model Selection: In this section, we discuss the selection and evaluation of the text summarization models that form the core of our project. we carefully selected and evaluated pre-trained text summarization models to generate concise and informative summaries from collected news articles. The primary model types considered were sequence-to-sequence (seq2seq) models and transformer models. Seq2seq models utilize an encoder-decoder architecture where the encoder processes the source text into a vector representation, which the decoder then uses to produce a summary. In contrast, transformer models leverage a neural network architecture capable of understanding word relationships within the text. Among the models evaluated, four stood out for their high performance in summarization tasks as mentioned in [4]: *moussaKam/barthez-orangesum-abstract*, a seq2seq model based on BART; *turner007/pegasus-summarizer*, a pretrained transformer model from the Pegasus family optimized for capturing essential text information; *facebook/bart-large*, another seq2seq model based on BART; and *gpt-3.5-turbo-16k*, a transformer-based model from the GPT-3 line, known for its advanced summarization capabilities. Each of these models was selected based on its ability to effectively transform extensive content into concise, meaningful summaries.

- Dataset and Model Evaluation: We conducted a comprehensive evaluation using standard metrics such as BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) to select the most suitable model for news article summarization. These metrics are widely recognized for assessing the quality of summaries generated by text summarization models, providing a quantitative measure of model performance. BLEU evaluates the precision of shared n-grams between the generated summary and human reference summaries, while ROUGE focuses on recall by measuring the overlap of words and phrases. Together, they offer a balanced perspective on the model's ability to capture the essence of the original content concisely and accurately. The evaluation utilized the CNN/Daily Mail dataset, a benchmark dataset widely used in summarization tasks. This dataset contains over 300,000 news articles from CNN and the Daily Mail, each paired with editor-crafted highlights that serve as high-quality reference summaries. Specifically, the dataset includes 286,817 training pairs, 13,368 validation pairs, and 11,487 test pairs, with an average article length of 766 words and summary length of 53 words. Such a structure is well-suited for testing the ability of models to generate concise and comprehensive summaries. We tested four summarization models on this dataset: GPT-3.5-turbo-16k, turner007/pegasus-summarizer, moussaKam/barthez-orangesum-abstract, and facebook/bart-large. Each model's summarization quality was objectively assessed using ROUGE-N, ROUGE-L, and BLEU metrics, measuring the overlap and sequence similarity between the model-generated and reference summaries. A subset of 1,000 articles from the test set was selected to ensure a thorough evaluation across various topics, including politics, business, and technology. The performance comparison of models and their analysis are shown in Table I.

Table I. Table denoting scores of different models for Summarization

S. No.	Models	Score BLEU (in%)	Score ROUGE (in %)
1.	Gpt-3.5-turbo-16k	16.39	66
2.	Turner007/Pegasus_summarizer	15.45	45
3.	moussaKam/barthez-orangesum-abstract	4.90	38
4.	Facebook/bart-large	5.38	29

Fig. 2 Performance comparison of different models for summarization

Among the models evaluated, the GPT-3.5-turbo-16k achieved the highest performance as shown in Table I with a BLEU score of 16.39% and a ROUGE score of 0.66, demonstrating superior alignment with human-crafted summaries. The turner007/pegasus-summarizer also showed strong performance, with a BLEU score of 15.45% and a ROUGE score of 0.45%, confirming its ability to generate coherent and concise summaries. While moussaKam/barthez-orangesum-abstract and facebook/bart-large provided valuable baseline results, they yielded lower BLEU and ROUGE scores, suggesting a relatively weaker ability to capture the content's key points with precision. These results, illustrated in Fig. 2, validate GPT-3.5-turbo-16k as the optimal model for our summarization task, given its robust ability to produce summaries that closely mirror human references.

3) Two-level Summarization with GPT 3.5

To utilize GPT-3.5-turbo-16k for summarization in our pipeline, we have integrated OpenAI's LangChain, which simplifies the process of API calls, providing enhanced control over requests, response handling, and batch processing. LangChain is particularly useful for managing complex prompts and handling text chunking, making it a more feasible solution than direct API calls to GPT-3.5. The API key for OpenAI is securely configured in the environment, allowing LangChain seamless access to GPT-3.5 functionalities.

We employ a two-level summarization approach to distill extensive news articles into concise, informative summaries, which is crucial given the token limits of the GPT-3.5-turbo-0125 model. This method ensures that we capture the complete content of each article without exceeding token constraints, enhancing efficiency and content coverage. The two-level summarization approach involves the following steps:

- Chunking of Text: To accommodate the token limit constraints of GPT-3.5, we utilize the Natural Language Toolkit (NLTK) to divide each article into smaller, manageable text segments. This chunking ensures that each segment adheres to the model's input context window limit of 16,385 tokens, and the summarized output for each segment remains within the 4,096-token maximum output limit [16]. This structured segmentation allows the model to process and summarize sections of content effectively without truncating important information.
- Initial Summarization Using GPT-3.5: After the text is divided into chunks, each segment is processed individually by the GPT-3.5-turbo-0125 model. The model generates a concise summary for each segment, capturing the essential points. This initial summarization ensures that even complex or lengthy articles are adequately represented in the first set of summarized outputs, preserving critical information.
- Refined Summarization and Final Aggregation: Following the initial summarization, the individual summaries from each chunk are further processed in a second-level summarization. The goal of this step is to combine and refine these segment-based summaries into a cohesive and comprehensive final summary. This second-level process is essential for maintaining coherence across the article's different sections and minimizing redundancy, thus providing a holistic overview that encapsulates the main points of the entire article without exceeding the model's token limitations.

This two-step summarization strategy enables effective content distillation by ensuring comprehensive coverage and eliminating the constraints posed by token limits. After completing the final aggregation, the unified summary is formatted as structured text and loaded into a Pandas DataFrame, preparing it for the subsequent sentiment analysis stage in the pipeline. This process ensures that users receive high-quality, accurate summaries that encapsulate the essential insights of each news article.

4) Sentimental Analysis

Sentiment analysis is a crucial aspect of understanding the emotional tone and bias in news content, providing insights into how news is presented and perceived. In our pipeline, sentiment analysis is performed after the summarization step to classify each news article as positive, negative, or neutral. This allows readers to gauge the overall sentiment of news stories, facilitating a more informed and personalized reading experience.

- Leveraging BERTweet Model: For the sentiment classification task, we selected the BERTweet model (finite automata/bertweet-base-sentiment-analysis), a BERT-based model fine-tuned specifically for sentiment tasks. BERTweet, pre-trained on a large corpus of Twitter data and optimized for sentiment detection, has proven highly effective in analyzing sentiment due to its bidirectional text processing. Unlike unidirectional models like GPT, BERTweet reads text both left-to-right and right-to-left, capturing deeper contextual nuances that enhance its ability to detect sentiment in complex sentence structures. This bidirectional approach makes BERTweet particularly suitable for sentiment analysis in news articles, where capturing context from both sides of a word or phrase is essential.

- Integration in the Pipeline: After each news article is summarized by the GPT-3.5 model, the summary is sent to the BERTweet model for sentiment analysis. Integrated into our pipeline via the Transformers library, BERTweet efficiently labels each summary as positive, negative, or neutral. These sentiment labels are then stored alongside the article summaries in a Pandas DataFrame, creating a structured format that links each news article with its respective sentiment. This process not only enhances the readability of the report but also offers a quick sentiment-based view of the news content, enabling users to better understand the tone of each article at a glance.
- Testing and Validation with the BBC News Dataset: To validate BERTweet's performance in a news context, we tested it on the BBC News dataset, which consists of 2,225 news articles spanning five topical categories: business, entertainment, politics, sports, and tech. The BBC News dataset provides a structured, labeled collection of articles across diverse domains, making it an ideal benchmark for testing BERTweet's sentiment classification abilities on news data. In this validation process, each article was categorized as positive, negative, or neutral based on its sentiment, allowing us to observe BERTweet's effectiveness in sentiment classification across various news topics as shown in Table II.

Table II. Table denoting sentiment results of the BERTweet model

S. No.	Topic/News Class	Total articles	Positive	Negative	Neutral
1.	Business	510	254	225	31
2.	Entertainment	401	163	210	28
3.	Politics	417	200	200	17
4.	Sports	511	236	246	29
5.	Technology	401	160	221	20

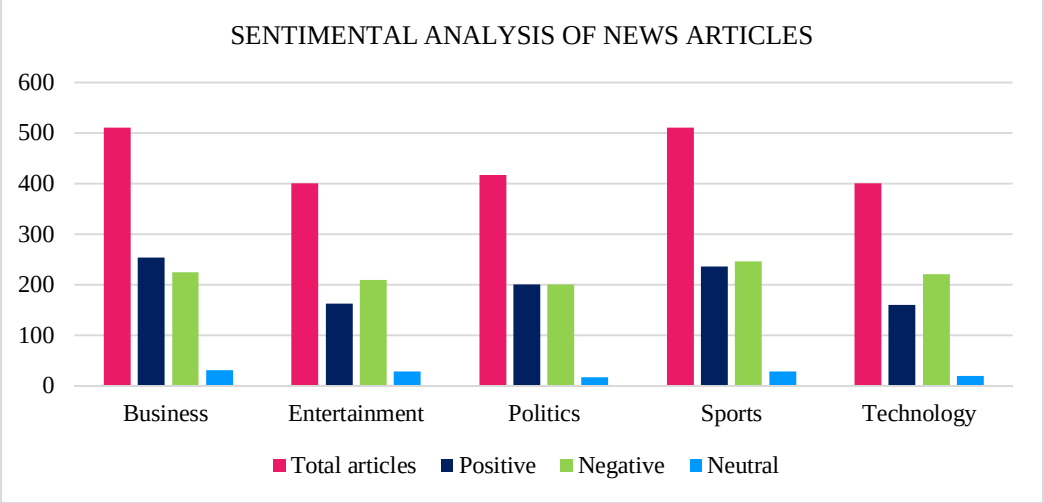


Fig. 3 Sentimental Analysis of News Articles across various categories

Fig. 3 presents the results of BERTweet's sentiment analysis on the BBC News dataset, revealing distinct sentiment patterns across various news categories. Most articles in the entertainment and tech categories exhibited negative sentiment, while business and sports articles were predominantly positive. The politics category displayed a balanced mix of positive and negative sentiments, reflecting a more nuanced tone. These findings, further illustrated in Table 2 and Fig. 2, demonstrate BERTweet's capacity to accurately classify sentiment within structured news content. BERTweet achieved an overall accuracy of approximately 89% on the BBC News dataset, with precision, recall, and F1-score values of 0.89, 0.88, and 0.89, respectively, across sentiment labels (positive, negative, and neutral). These metrics highlight BERTweet's suitability for large-scale, automated sentiment analysis applications, and ability to perform accurate sentiment classification on structured news content.

- Sentiment Labeling: After sentiment analysis, each article's sentiment label (positive, or negative, or neutral) is appended to the Pandas DataFrame This sentiment-labeled DataFrame forms the basis of the final report, which provides users with a sentiment-based overview of the news articles, allowing for easy and efficient access to news stories filtered by sentiment. The integration of BERTweet for sentiment analysis not only improves the quality of content delivery but also empowers users to engage with news articles based on their preferred emotional tone, enhancing the overall reading experience.

5) Report Generation

The final step in our project involves generating a well-organized report from the processed news data, providing users with a summarized, sentiment-labeled overview of the latest news articles. This report is structured in a Pandas DataFrame, where each row represents an article with columns for Title, Media Source, Timestamp, Description, Link, and Sentiment Label (positive, negative, or neutral). Once the data is organized, it is converted into HTML format using Pandas' to_html() function. This HTML conversion creates a clean, tabular format that is easy to embed directly in the body of an email, ensuring that recipients receive a structured, visually appealing summary. With this setup, users can conveniently access the latest news summary directly in their authenticated email, viewing each article's key details and sentiment label.

D. Gmail API Configuration and Email Delivery

To securely and reliably send the generated reports, we configure the Gmail API. Compared to traditional SMTP, the Gmail API provides enhanced security through OAuth 2.0 authentication and better scalability, making it suitable for sending larger volumes of email. The configuration process begins by registering the application in the Google Cloud Console and enabling the Gmail API, followed by creating OAuth 2.0 credentials for secure access. Once configured, the report, formatted as an HTML email is sent to the user's email address that was used for authentication. This approach ensures that the summary reaches users' verified email addresses, maintaining both security and consistency in report distribution.

The Gmail API's open-source nature and built-in security advantages make it preferable over SMTP where reliable and frequent email delivery is required. It also provides a robust infrastructure to automate the daily report process.

IV. FLOWCHART

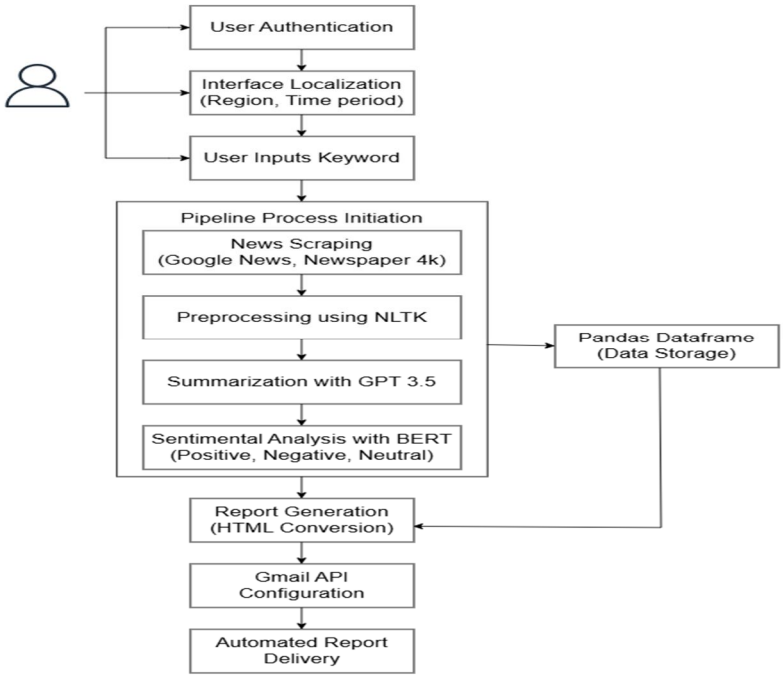


Fig. 4 Workflow of the proposed system

Fig. 4 presents a flowchart illustrating the workflow of processes involved in delivering the automated news report, from fetching news articles based on the user's keyword to final report generation.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XI Nov 2024- Available at www.ijraset.com

V. RESULT

The final report email contains a structured table with the news articles, including Title, Media Source, Timestamp, Description, Link, and Sentiment Label. The email's subject line, "Automated Daily Report," gives recipients an understanding of the report's purpose. The body includes a header summarizing the keyword focus, such as "Summary of GPT (Keyword) today: Summarization." Below this, the structured HTML table provides detailed insight into the latest news, complete with sentiment labels. The image of the email report in the inbox below illustrates the final product, showcasing how users receive and engage with the summarized news and sentiment analysis results.

Automated Daily Report
1 message

<deppjonney9@gmail.com> Wed, 8 May, 2024 at 11:53 am
To: dchandru1861@gmail.com

Summary of GPT today: The news content discusses the development of a new GPT-4-based AI chatbot by Microsoft for use by US intelligence agencies. This chatbot is designed to analyze classified information securely, as it is disconnected from the internet. The article also touches on the use of AI in government and military applications, as well as the concerns and advancements in AI governance and security.

Title	Media	Timestamp	Description	Link	sentiment_label
GPT-4 now has vision—can it actually read chest X-rays?	Health Imaging	2024-05-08 06:07:12.439800	Finely tuned, pre-trained large language models are beginning to reliably translate image content into text, but are they ready to take on medical images?10 hours ago	https://healthimaging.com/topics/artificial-intelligence/gpt-4-now-has-vision-can-it-actually-read-chest-x-rays&ved=2ahUKEwiYI7vIsf2FAxU6e_UHHTZVBscQxfQBegQIARAC&usg=AOvVaw02OUkDc7QExkQHc-UDSoit	NEU
BiSpec Pairwise AI: guiding the selection of bispecific antibody target combinations with pairwise learning and GPT ...	springermedizin.de	2024-05-08 06:00:07.209263	Malignant tumors represent a major global health challenge, often culminating in mortality due to the limitations of conventional treatments like ...57 minutes ago	https://www.springermedizin.de/bispec-pairwise-ai-guiding-the-selection-of-bispecific-antibody-/27064916&ved=2ahUKEwiYI7vIsf2FAxU6e_UHHTZVBscQxfQBegQIBBAC&usg=AOvVaw0puvqmQ2w5jBrDbc1REtqZ	NEU
In 12 Months, OpenAI Predicts ChatGPT Will Outshine Current Version, Rendering It 'Laughably Bad'	TechStory	2024-05-08 05:07:12.445291	During the 27th annual Milken Institute Global Conference, the atmosphere buzzed with anticipation as OpenAI's COO Brad Lightcap took the stage.	https://techstory.in/in-12-months-openai-predicts-chatgpt-will-outshine-current-version-rendering-it-laughably-bad/&ved=2ahUKEwiYI7vIsf2FAxU6e_UHHTZVBscQxfQBegQIBRAC&usg=AOvVaw22hLWEQrrR2WMEW5AUhbVa	NEG
Microsoft launches AI chatbot for spies	Ars Technica	2024-05-08 05:07:12.442675	Air-gapping GPT-4 model on secure network won't prevent it from potentially making things up.	https://arstechnica.com/information-technology/2024/05/microsoft-launches-ai-chatbot-for-spies/&ved=2ahUKEwiYI7vIsf2FAxU6e_UHHTZVBscQxfQBegQIABAC&usg=AOvVaw3VerlikpkSK6L17QuWuKLy	NEG

Fig. 5 Report received in the user's inbox

Title	Media	Timestamp	Description	Link	sentiment_label
Artificial Intelligence and GPT-Powered Skin Epigenetics: Applications in Personalized Skincare | What is Epigenetics?	WhatIsEpigenetics.com	2024-05-08 02:07:12.452883	Epigenetics refers to the modifications of DNA by adding chemical tags, which switch genes on and off without altering the underlying DNA sequence.	https://www.whatisepigenetics.com/artificial-intelligence-and-gpt-powered-skin-epigenetics-applications-in-personalized-skincare/&ved=2ahUKEwiYI7vIsf2FAxU6e_UHHTZVBscQxfQBegQIAxAC&usg=AOvVaw1tSAF7LU4avaJvgFMy6wK4	NEU
GPT-Based AI Chatbot for Spies: Did Microsoft Develop One for US Intelligence?	Tech Times	2024-05-08 02:07:12.312065	A new and secure GPT-4-based AI chatbot is here from Microsoft, and it is designed by the company for spies and US intelligence to use for its needs.	https://www.techtimes.com/articles/304406/20240507/gpt-based-ai-chatbot-spies-microsoft-made-one-intelligence.htm&ved=2ahUKEwiYI7vIsf2FAxU6e_UHHTZVBscQxfQBegQICRAC&usg=AOvVaw1MeyNBI2dj2Doajuc5WD4k	NEU
Stardock's DesktopGPT now lets you use GPT-4 Turbo without the browser	XDA Developers	2024-05-08 01:07:12.875761	Want a GPT-4 Turbo chatbot on your desktop? Stardock has you sorted.	https://www.xda-developers.com/stardock-desktopgpt-gpt-4-turbo/&ved=2ahUKEwj_8rfosf2FAxUebvUHHQJmBnA4ChDF9AF6BAgJEAI&usg=AOvVaw3sLpBLQDQ66dN6ulE0gvjk	NEG
AI Risk launches GPT for RIAs	Citywire	2024-05-08 01:07:12.872770	The GPT will provide AI agents including 'personal assistant,' 'sales consultant' and 'marketing strategist,' as well as cybersecurity and compliance...	https://citywire.com/ria/news/ai-risk-launches-gpt-for-rias/a2441873&ved=2ahUKEwj_8rfosf2FAxUebvUHHQJmBnA4ChDF9AF6BAgFEAI&usg=AOvVaw3SlhdrdUMNVTNBWfJ1sYTE	NEU
GPT Group Announces March 2024 Progress - TipRanks.com	Tipranks	2024-05-08 01:07:12.347919	GPT Group (AU:GPT) has released an update. The GPT Group has released its March 2024 quarterly update, marking another period of progress for the company.	https://www.tipranks.com/news/company-announcements/gpt-group-announces-march-2024-progress&ved=2ahUKEwiYI7vIsf2FAxU6e_UHHTZVBscQxfQBegQIAhAC&usg=AOvVaw20p0FBUyxLjIDd3PFCkM74	NEU

Fig. 6 Continuation of the report received in the user's inbox

VI. FUTURE WORK

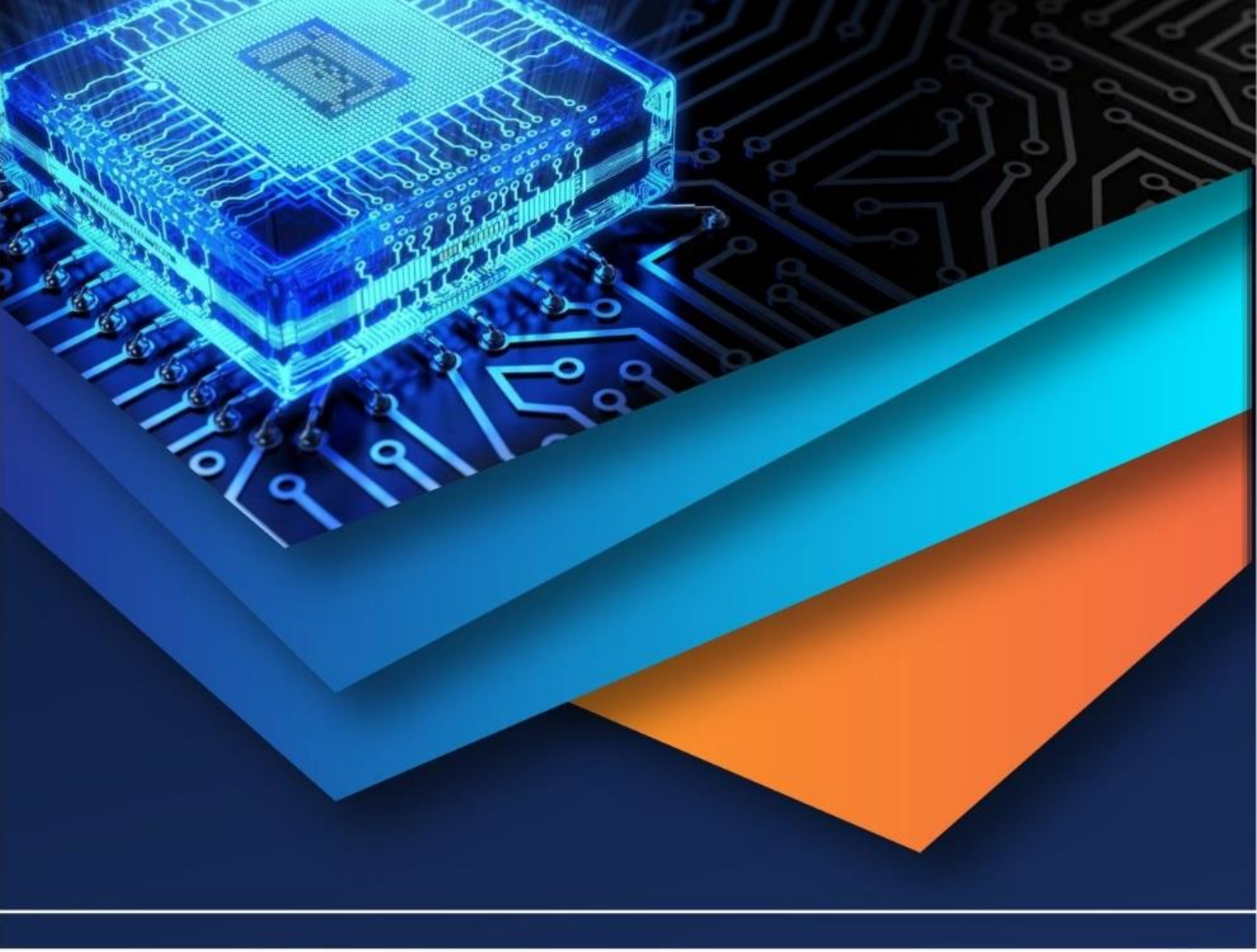
Future iterations of this project will focus on multilingual capabilities, enabling users to read summaries in their preferred language [15]. By incorporating language detection and transfer learning-based translation models, we can expand access to a broader range of global news sources. Leveraging APIs such as Google News and Bing News for multilingual articles, the system will dynamically translate content into English (or another selected language) for summarization. Translation APIs like Google Translate or DeepL, or a custom fine-tuned translation model, will enable consistent and accurate translations. This approach aims to create an inclusive platform, catering to diverse linguistic audiences while offering reliable news summaries in users' preferred languages.

VII. CONCLUSION

This research illustrates the effectiveness of utilizing NLP models, such as GPT-3.5 and BERTweet, in automating daily news summarization and sentiment analysis. By implementing a fully integrated pipeline for news scraping, summarization, and sentiment classification, the system delivers concise, sentiment-rich reports directly to users' emails. With a scalable architecture and advanced model integrations, this approach enhances user accessibility and engagement with news content. Expanding this system to incorporate multilingual capabilities, along with user-specific localization, will allow for a more inclusive platform, delivering relevant news summaries in preferred languages. Our application thus holds the potential to transform real-time news analysis, providing users with a comprehensive yet personalized perspective on current events.

REFERENCES

[1] H. Batra, A. Jain, G. Bisht, K. Srivastava, M. Bharadwaj, D. Bajaj, and U. Bharti, "News Summarization Application Based on Deep NLP Transformers for SARS-CoV-2," Proc. Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1–6, 2021, DOI: 10.1109/ICRITO51393.2021.9596520.
[2] S. Tarannum, P. Sonar, A. Agrawal, and K. Khairnar, "NLP-based Text Summarization Techniques for News Articles: Approaches and Challenges," Int. Res. J. Eng. Technol. (IRJET), vol. 8, no. 12, pp. 500–508, 2021.
[3] H. Jin, Y. Zhang, D. Meng, J. Wang, and J. Tan, "A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods," IEEE Trans. Knowl. Data Eng., pp. 1–21, 2024, arXiv:2403.02901v1.
[4] I. F. Compaoré, R. Kafando, A. K. Kabore, S. Aminata, and T. F. Bissyandé, "AI-driven Generation of News Summaries: Leveraging GPT and Pegasus Summarizer for Efficient Information Extraction," HAL Open Archive, ffhal-04536137f, 2024.
[5] J. O. Krugmann and J. Hartmann, "Sentiment Analysis in the Age of Generative AI," Customer Needs and Solutions, vol. 11, no. 3, pp. 145–160, 2024, DOI: 10.1007/s40547-024-00143-4.
[6] B. Khan, Z. A. Shah, M. Usman, I. Khan, and B. Niazi, "Exploring the Landscape of Automatic Text Summarization: A Comprehensive Survey," IEEE Access, vol. 11, pp. 1–30, 2023, DOI: 10.1109/ACCESS.2023.3322188.
[7] M. A. Mutlu, E. E. Ulku, and K. Yildiz, "A Web Scraping App for Smart Literature Search of Keywords," PeerJ Comput. Sci., vol. 10, pp. 1–12, 2024, DOI: 10.7717/peerj-cs.2384.
[8] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent Advancements and Challenges of NLP-Based Sentiment Analysis: A State-of-the-Art Review," Nat. Lang. Process. J., vol. 6, p. 100059, 2024, DOI: 10.1016/j.nlp.2024.100059.
[9] A. B. Salunke, E. Saini, S. Shinde, P. Tumma, and S. S. Dange, "NewsIN: A News Summarizer and Analyzer," Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET), vol. 10, no. 12, pp. 800–810, 2022.
[10] N. B. Ahmad, "Application of Pre-trained Models (PTMs) in Sentiment Analysis, News Classification, Anti-Spam Detection, and Information Extraction," J. Comput. Soc. Dyn., vol. 8, pp. 1–15, 2023.
[11] J. Ahmed and M. Ahmed, "A Framework for Sentiment Analysis of Online News Articles," J. Comput. Res., vol. 7, pp. 100–110, 2020.
[12] K. Pipalia, R. Bhadja, and M. Shukla, "Comparative Analysis of Different Transformer Based Architectures Used in Sentiment Analysis," Proc. 9th Int. Conf. Syst. Model. Adv. Res. Trends (SMART), pp. 50–55, 2020, DOI: 10.1109/SMART50582.2020.9337081.
[13] S. T. Kokab, S. Asghar, and S. Naz, "Transformer-based Deep Learning Models for the Sentiment Analysis of Social Media Data," Array, vol. 14, p. 100157, 2022, DOI: 10.1016/j.array.2022.100157.
[14] V. Deokar and K. Shah, "Automated Text Summarization of News Articles," Int. Res. J. Eng. Technol. (IRJET), vol. 8, no. 9, pp. 400–410, 2021.
[15] R. G. Goriparthi, "AI-Driven Natural Language Processing for Multilingual Text Summarization and Translation," Rev. Intell. Artif. Med., vol. 12, no. 1, pp. 1–10, 2021.
[16] OpenAI, "GPT-3.5 Model Overview and Documentation from OpenAI," OpenAI Documentation, 2024. [Online]. Available: https://platform.openai.com/docs/models.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY