Project Report

Media Monitoring

Multi-label Classification

Team : Binary Brigade

Prepared by

Prkhar Mishra
(IIT Indore)

Aditi Chaturvedi
(Chandigarh University)

Table of Contents

1. INTRODUCTION

1.1 Overview

In an increasingly digitized world, where data and information proliferate at an unprecedented rate, efficient management, classification, and analysis of media content have become paramount. The "Media Monitoring Multilabel Classification" project addresses the burgeoning need of media monitoring companies to sift through vast volumes of printed media articles from newspapers, magazines, and online sources to extract valuable insights.

Traditionally, this process was a labor-intensive endeavor, relying heavily on manual classification by human analysts. However, with the advent of machine learning and natural language processing (NLP) techniques, the landscape of media monitoring has been transformed. This project harnesses the power of these technologies to develop a state-of-the-art multi-label classification system capable of accurately categorizing printed media articles into multiple relevant topics.

1.2 Purpose

The core purpose of this project is to create an automated solution that not only enhances the efficiency of media monitoring operations but also significantly improves the effectiveness of content classification. By automating the classification process, media monitoring companies can rapidly process and categorize articles, allowing them to extract valuable insights and trends more promptly than ever before.

The project's significance lies in its potential to streamline the workflow of media monitoring organizations, reduce manual labor costs, and increase the speed and accuracy of content analysis. Furthermore, the system's ability to classify articles into multiple relevant topics provides a nuanced understanding of media content, enabling better-informed decision-making and trend analysis.

Key Objectives

The primary objectives of the "Media Monitoring Multilabel Classification" project include:

- Develop a data-driven multi-label media article classification system.
- Efficiently categorize printed media articles into multiple relevant topics.
- Utilize machine learning techniques, including Natural Language Processing (NLP) and image analysis, to automate the classification process.
- Enhance efficiency and effectiveness for media monitoring companies.
- Provide a user-friendly web interface for seamless interaction with the system.

The project's core focus is to bridge the gap between the growing volume of media content and the need for efficient analysis, enabling organizations to harness the power of data-driven insights from printed media articles. Through this automation, we aim to empower media monitoring companies to stay agile and responsive in a rapidly evolving media landscape.

2. LITERATURE SURVEY

2.1 Existing Problem

In the realm of media monitoring and content analysis, several persistent challenges have been identified. These challenges highlight the need for innovative solutions such as the "Media Monitoring Multilabel Classification" project:

- Information Overload: The digital age has led to an explosion of media content across various platforms. Media monitoring companies are overwhelmed by the sheer volume of information they need to analyze daily. Human analysts struggle to keep pace with the continuous stream of news articles, resulting in potential delays in insights extraction.
- Manual Classification: Traditional media monitoring heavily relies on manual classification by human analysts. This approach is not only time-consuming but also prone to errors and inconsistencies in categorizing articles into relevant topics.
- Scalability Issues: As media monitoring companies expand their services to cover a broader range of topics and sources, scalability becomes a critical issue. Manual processes are inherently limited in their ability to handle large volumes of data efficiently.
- Lack of Nuance: Single-label classification systems often fall short in capturing the nuanced nature of media content. Articles frequently cover multiple topics simultaneously, making it challenging to assign a single category accurately.
- Incomplete Insights: Manual analysis may result in partial or incomplete insights extraction, as human analysts may overlook crucial information or trends due to time constraints and the overwhelming amount of content.

FIGURE 1: Existing Approach Process

2.2 Proposed Solution

The core purpose of this project is to create an automated solution that not only enhances the efficiency of media monitoring operations but also significantly improves the effectiveness of content classification. By automating the classification process, media monitoring companies can rapidly process and categorize articles, allowing them to extract valuable insights and trends more promptly than ever before.The "Media Monitoring Multilabel Classification" project proposes an innovative solution to address the aforementioned challenges:

- Automated Classification: The core of the solution lies in the development of a data-driven, multi-label classification system. This system leverages advanced machine learning techniques, including Natural Language Processing (NLP) and image analysis, to automate the classification process of printed media articles.

- Enhanced Efficiency: By automating classification, the project significantly enhances the efficiency of media monitoring operations. The system can rapidly process and categorize articles, reducing the time and effort required for manual analysis.

- Multi-Label Classification: Unlike traditional single-label classifiers, the proposed system can classify articles into multiple relevant topics simultaneously. This nuanced approach ensures a more accurate representation of article content, capturing the complexity of real-world media narratives using IBM Watson studio.

- Insights Extraction: Through automated classification, sentiment analysis, entity recognition, and summarization, the system provides a comprehensive framework for extracting valuable insights from media content. This enables media monitoring companies to make well-informed decisions and identify emerging trends swiftly.

- Scalability and Speed: The project's automated approach is inherently scalable, enabling media monitoring companies to process a growing volume of articles without sacrificing accuracy or speed.

In conclusion, the "Media Monitoring Multilabel Classification" project seeks to revolutionize the media monitoring industry by offering a modern solution to the enduring challenges of information overload, manual classification, and incomplete insights extraction. Through the integration of cutting-edge technologies, this project aims to empower media monitoring companies to operate efficiently, make data-driven decisions, and stay competitive in the fast-paced world of media analysis.

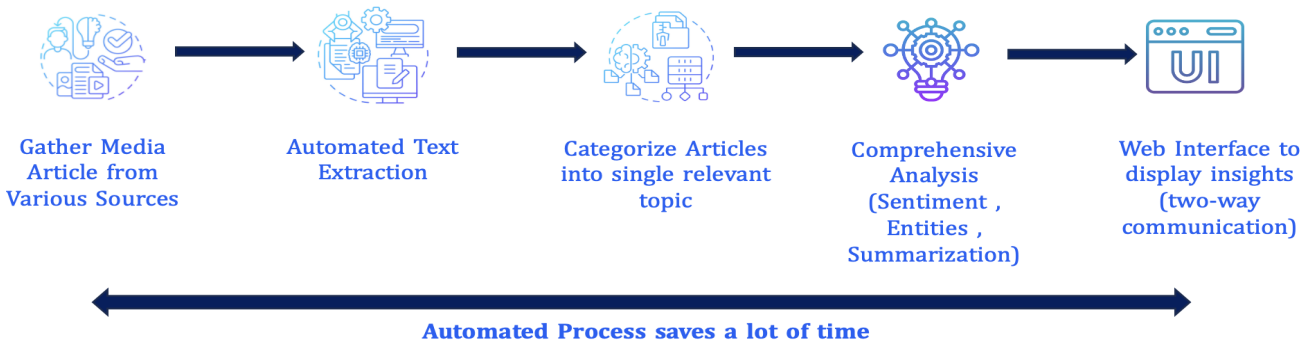


FIGURE 2: Proposed Solution Process

3. THEORITICAL ANALYSIS

3.1 Block Diagram

A diagrammatic overview of the project typically provides a high-level visual representation of the key components and their interactions within the entire project. It aims to give a holistic understanding of the project's structure and major elements. This type of diagram is often used in project proposals, presentations, or introductory sections of project reports. It doesn't focus on the detailed control flow but rather on the project's architecture. Summary of the steps outlined in the block diagram:

- Gather Media Articles: Obtain media articles from diverse sources, including web, files, and more.
- Automated Text Extraction: Employ OpenCV for image-to-text conversion and BeautifulSoup for web scraping to extract article content.
- Data Pre-processing: Prepare and clean the data by handling tasks like text cleaning and tokenization.
- Multi-Label Classification: Utilize advanced machine learning techniques like BERT and traditional ML models for accurate topic classification using IBM Watson studio.
- Sentiment Analysis: Apply Natural Language Processing tools to analyze and understand the emotional context within articles.
- Entity Recognition: Implement BERT Tokenizer to identify organizations, events, and named entities.

- Summarization: Use BERT Summarizer and text summarization methods to generate concise article summaries.

- Web Interface to display insights: Develop a user interface with Django and React for user interaction. Showcase the processed data, including classifications, sentiment, and summaries, in the frontend UI.

These steps collectively form a comprehensive Media Monitoring Multilabel Classification System, automating the analysis of media articles, and enhancing efficiency and effectiveness in media monitoring tasks.

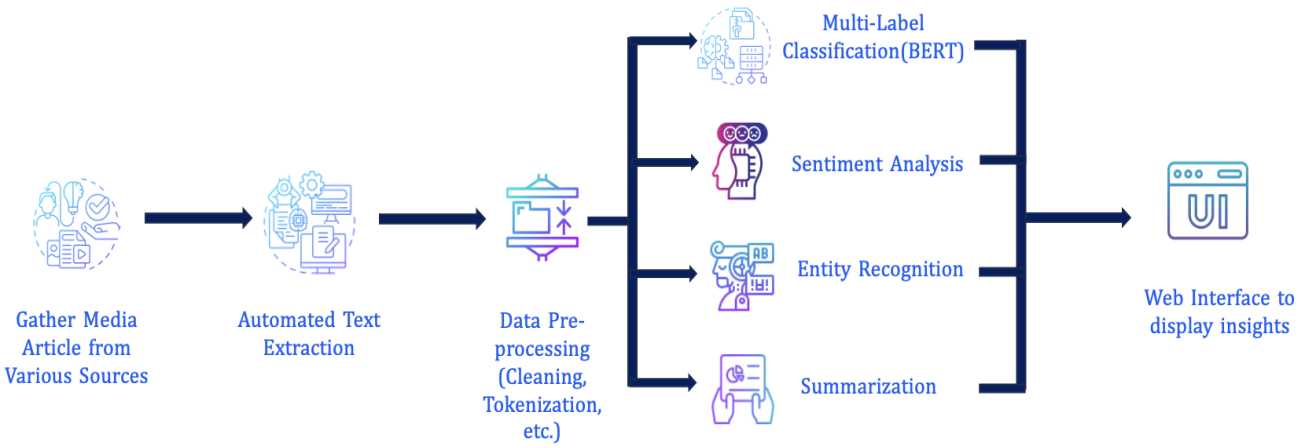


FIGURE 3: Diagrammatic Overview

3.2 Hardware / Software Designing

Hardware Requirements

The successful implementation of the Media Monitoring Multilabel Classification System necessitates certain hardware components to support its functionality. These components are fundamental to enabling the efficient execution of various tasks involved in the project. The key hardware requirements include:

- Computer Systems: High-performance computing systems capable of handling machine learning tasks and data processing efficiently.

- Storage: Adequate storage space to accommodate the dataset, models, and intermediary results generated during the project.

- GPU (Graphics Processing Unit): For accelerated machine learning model training, especially when using deep learning techniques such as BERT.
- Internet Connectivity: A stable and high-speed internet connection is essential for web scraping and data retrieval from online sources.
- Web Server: If deploying the project on a web server for production use, server hardware and resources will be required.

Software Requirements

The software requirements for the Media Monitoring Multilabel Classification System are diverse and span various layers of the solution. These software components provide the necessary tools and frameworks to implement the project successfully. The key software requirements include:

- Operating System: The project can be developed and executed on multiple operating systems, including Windows, Linux, or macOS.
- Python: The primary programming language for implementing machine learning models and data processing.
- Integrated Development Environment (IDE): Popular choices include Jupyter Notebook, PyCharm, or Visual Studio Code for Python development.
- Libraries and Frameworks: Several Python libraries and frameworks are integral to different aspects of the project:
- OpenCV: For image processing and extraction.
- BeautifulSoup: For web scraping and parsing HTML content.
- Natural Language Processing (NLP) Libraries: Such as NLTK or spaCy for text processing.
- Django: For developing the REST API.
- React: For building the user interface.
- Web Scraping Tools: Such as Scrapy for more advanced web scraping tasks.
- Machine Learning Frameworks: TensorFlow, PyTorch, and Hugging Face Transformers library for BERT model deployment and usage.
- Version Control: Git for version control and collaborative development.

- Containerization: Docker for containerizing the application and its dependencies.
- Web Hosting (Optional): If deploying the project to the web, a hosting service IBM Cloud

These hardware and software components collectively enable the successful development, deployment, and operation of the Media Monitoring Multilabel Classification System.

4. Experimental Investigation

During the course of developing the Media Monitoring Multilabel Classification System, a series of rigorous experimental investigations were conducted to assess the effectiveness and performance of the various components and algorithms employed. These experiments were essential for fine-tuning the system, evaluating its capabilities, and ensuring that it meets the project objectives. The investigations can be categorized into several key areas:

1. Data Collection and Pre-processing

Objective: To ensure that the data collected and processed for training and testing the models are of high quality and suitable for the intended tasks.

- Data Collection: A diverse dataset of printed media articles from newspapers, magazines, and online sources was gathered. This dataset consisted of articles labeled under specific categories like business, entertainment, politics, sport, or tech.
- Data Pre-processing: Various pre-processing steps were applied, including noise removal, handling missing values, and standardizing the format of the articles. Tokenization, stopword removal, stemming/lemmatization, and feature embeddings were employed to represent text content accurately. Image analysis techniques using OpenCV were used for meaningful feature extraction from images.

2. Model Selection and Training

Objective: To identify the most suitable machine learning algorithms and deep learning models for multi-label classification, sentiment analysis, entity recognition, and summarization.

- Exploration of Models: Multiple machine learning algorithms, including Support Vector Machines (SVM), Random Forests, and deep learning models like Convolutional Neural

Networks (CNNs), Recurrent Neural Networks (RNNs), and BERT, were considered.

- Training and Hyperparameter Tuning: The selected models were trained on the annotated dataset. Extensive hyperparameter tuning was performed to optimize their performance in terms of accuracy, precision, recall, and F1-score.

3. Model Evaluation

Objective: To assess the performance of the trained models using various metrics.

- Performance Metrics: The models' performance was evaluated using metrics such as precision, recall, F1-score, and accuracy. These metrics were calculated for multi-label classification, sentiment analysis, entity recognition, and summarization tasks.

4. Web Interface Development

Objective: To create a user-friendly web interface for users to interact with the system.

- User Interface Testing: The developed web interface was rigorously tested to ensure smooth user interaction and functionality. User feedback was collected and incorporated for improvements.

5. Data-Driven Classification

Objective: To validate the system's ability to classify media articles accurately into relevant topics.

- Media Article Testing: A set of media articles, including text, images, and URLs, was submitted through the web interface. The system processed the content to extract relevant information and employed the trained machine learning models to classify articles into multiple relevant topics based on extracted features.

5. FLOWCHART

This sequential diagram illustrates the flow of tasks in the Media Monitoring Multilabel This sequential diagram outlines the step-by-step process of the Media Monitoring Multilabel Classification System. It begins with the User Interface Layer, where users provide input in the form of text, images, or URLs. The Logical Layer then takes over, starting with input type detection. Depending on whether the input is text, an image, or a URL, specific processing paths are followed. For text input, the system proceeds to employ a BERT Classifier for text classification, assigning multiple tags or categories to the content. Additionally, sentiment analysis is performed to gauge the emotional tone of the input. For images, the system utilizes OpenCV for image extraction, and for URLs, web scraping with BeautifulSoup is conducted to obtain relevant content. Following these initial steps, the Logical Layer employs a BERT Tokenizer to extract entities like organizations and events from the text, and a BERT Summarizer is used to generate concise summaries of news articles.

The results and processed data are then presented to users through the User Interface Layer, which is built using React. Users can see classification tags, sentiment analysis results, extracted entities, and summarized news articles. This sequential flow ensures that user-provided input, regardless of its form, undergoes thorough analysis and processing, ultimately delivering valuable insights and information in a user-friendly manner. The integration of various layers and components enables efficient interaction and information retrieval, making the system a powerful tool for media monitoring and multilabel classification.

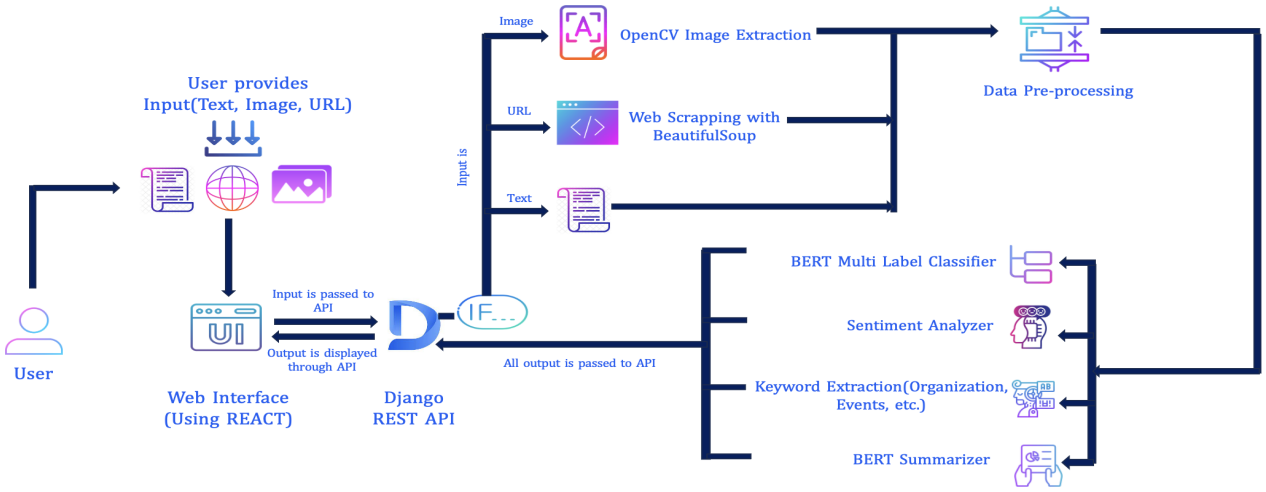
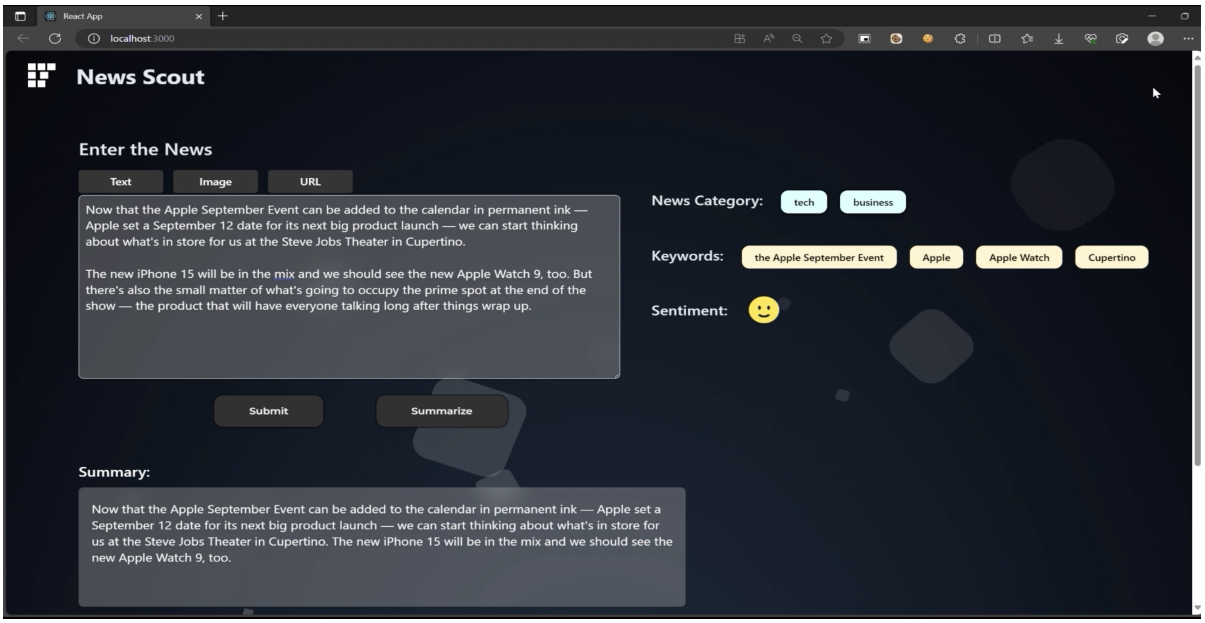


FIGURE 4: Control flow of solution

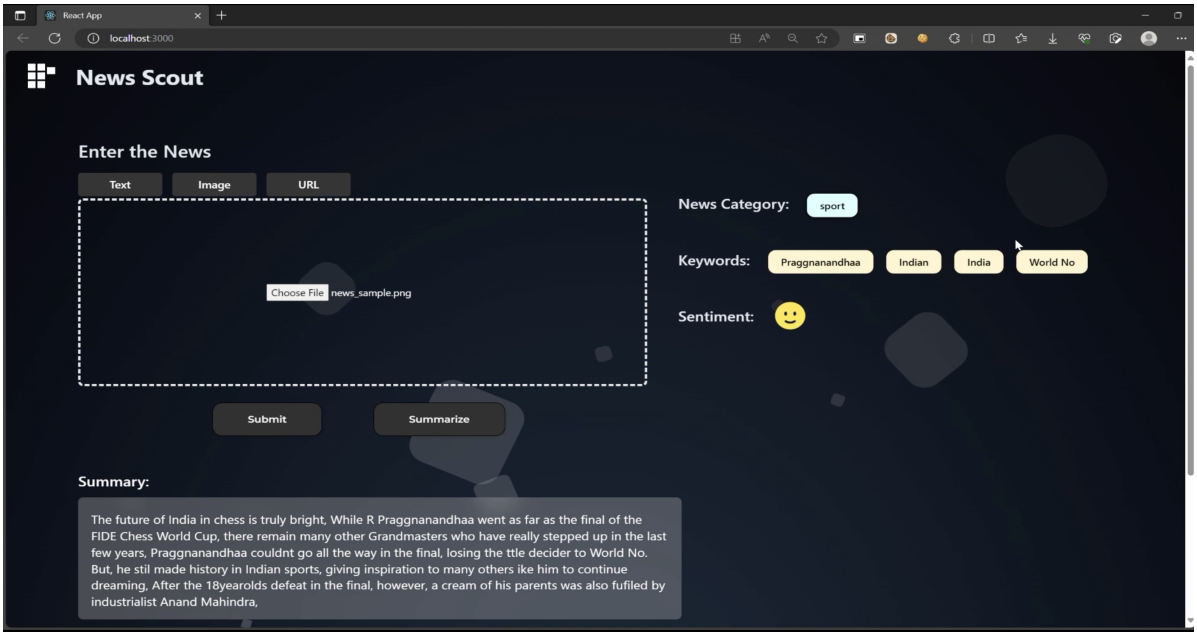
6. RESULT

Here are the News Category Tagging , Keywords, Sentiments and Summarization of media using following Inputs.

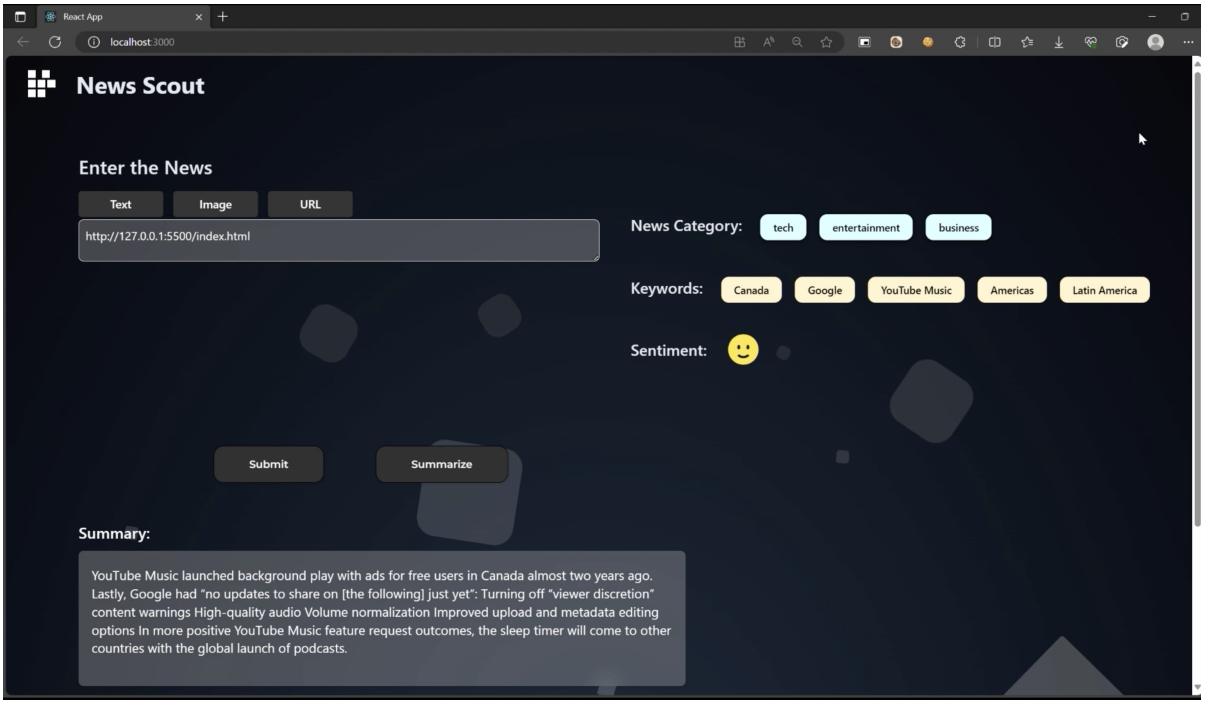
1. **TEXT**



2. **IMAGE**



3. **URL**



7. ADVANTAGES & DISADVANTAGES

Advantages:

- Efficient Classification: The system uses advanced machine learning techniques, including BERT-based models, for precise and efficient multilabel classification of media articles, reducing manual effort.
- Multi-Input Support: It accommodates text, images, and URLs as input, making it versatile for handling diverse data sources and formats.
- Comprehensive Analysis: The system includes sentiment analysis, entity recognition, and summarization, providing valuable insights beyond categorization.
- User-Friendly Interface: Developed with React, the interface is intuitive for users, enabling easy input and access to categorized tags, sentiment analysis, entity extraction, and summarized content.
- Scalability: With machine learning models and a REST API, it's scalable to handle increased data volumes and user loads.

Disadvantages:

- Data Quality Dependency: The system's effectiveness relies on high-quality training data; inaccurate data may lead to misclassifications.
- Resource Intensive: Training and running deep learning models, like BERT, can be resource-demanding, posing challenges for resource-constrained environments.
- Complexity: Integrating various components, including BERT, introduces complexity to system setup and maintenance, requiring skilled personnel.
- Language Limitation: Primarily tailored for English text, the system's effectiveness may diminish with non-English content.
- Web Scraping Challenges: Web scraping with BeautifulSoup may face hurdles due to evolving website structures or restrictions, necessitating vigilance.
- False Positives/Negatives: Automated systems can produce false positives (incorrect labels) and false negatives (missed labels), requiring ongoing refinement.
- Privacy Concerns: Processing user-submitted URLs may raise privacy concerns, necessitating robust data handling and protection.

In conclusion, the Media Monitoring Multilabel Classification System automates media monitoring but faces challenges like data quality and resource demands that need careful management.

8. APPLICATIONS

The Media Monitoring Multilabel Classification System offers a wide array of practical applications across various domains. Its ability to efficiently process and classify printed media articles into multiple relevant topics, along with additional features like sentiment analysis and entity extraction, makes it a versatile tool. Here are some key areas where this solution can be applied:

- **Media Analysis:** Media monitoring companies and news agencies can streamline news categorization and trend analysis, facilitating comprehensive media coverage assessment.
- **Market Research:** Market analysts can efficiently track industry trends, sentiment, and competitor activities through categorized media content.
- **Public Relations:** PR professionals use the system to monitor media presence, assess public sentiment, and address potential PR issues.
- **Competitive Intelligence:** Businesses gain insights into competitors by categorizing articles, enabling informed strategic decisions.
- **Government and Policy Analysis:** Government agencies gauge public sentiment on policy topics and identify areas for public information campaigns.
- **Brand Management:** Brands monitor media mentions, assess sentiment, and proactively manage their reputation.
- **Research and Academia:** Researchers employ the system for literature review, topic analysis, and sentiment assessment in academic studies.
- **Content Recommendation:** Online platforms improve content recommendations by categorizing and summarizing articles for personalized user experiences.
- **Compliance and Regulation:** Industries subject to regulations ensure compliance and monitor media coverage, e.g., healthcare and finance.
- **Social Listening:** Social media monitoring extends to traditional media, offering a comprehensive view of public sentiment.

The Media Monitoring Multilabel Classification System adapts to diverse industries, providing timely and accurate media insights for informed decision-making and strategic planning.

9. CONCLUSION

In the realm of media monitoring and multilabel classification, the journey embarked upon with the creation of the Media Monitoring Multilabel Classification System has been both enlightening and transformative. This endeavor sought to address the challenges inherent in the analysis of printed media articles, leveraging cutting-edge machine learning techniques and tools to automate and enhance the efficiency and effectiveness of the classification process.

Throughout this project, we have achieved significant milestones and garnered valuable insights:

Automation and Efficiency: The heart of our system lies in its ability to automate the classification of printed media articles into multiple relevant topics. By harnessing the power of Natural Language Processing (NLP), image analysis, and deep learning models such as BERT, we have reduced the arduous manual classification efforts that media monitoring companies once faced.

Comprehensive Insights: Our system goes beyond mere categorization. It delves into the sentiments expressed in articles, extracts essential entities like organizations and events, and provides concise article summaries. This comprehensive approach has illuminated the potential for richer, more nuanced media analysis.

Versatility and Adaptability: One of the system's hallmarks is its flexibility. It seamlessly handles diverse input types, including text, images, and URLs, making it adaptable to various data sources and formats. This adaptability extends to the range of applications, from media analysis to market research and brand management.

User-Centric Design: At the forefront of our system is a user-friendly interface developed with React. This interface ensures that users can effortlessly input media articles and receive insights and categorizations in a straightforward and intuitive manner.

As we reflect on this journey, it is evident that the Media Monitoring Multilabel Classification System represents a leap forward in the realm of media analysis. Its potential to drive efficiency, provide comprehensive insights, and cater to a multitude of applications makes it a valuable asset across industries and domains.

However, like any technological innovation, our system is not without its challenges. Data

quality, resource demands, and the evolving nature of web scraping are among the hurdles we've encountered. Nevertheless, these challenges are stepping stones toward continued improvement and refinement.

Looking ahead, the future of this system holds promise. With ongoing model enhancements, wider adoption, and adaptations for different languages, it will continue to evolve and meet the dynamic needs of media monitoring and analysis.

In closing, the Media Monitoring Multilabel Classification System represents not just a culmination of work, but a beginning—a beginning of a new era in media monitoring, where automation, efficiency, and comprehensive analysis converge to provide deeper insights and greater value to those who seek to understand the ever-evolving landscape of printed media.

10. FUTURE SCOPE

As we look to the future of the Media Monitoring Multilabel Classification System, several promising enhancements and developments await:

- Multilingual Proficiency: Expanding language support beyond English to encompass a multitude of languages will unlock global applicability.
- Real-Time Insights: Enabling real-time analysis ensures users receive up-to-the-minute categorizations, vital for industries where timeliness is paramount.
- Continuous Learning: Self-learning mechanisms will enable adaptation to evolving languages, trends, and topics, ensuring relevance and accuracy.
- Advanced Entity Recognition: Augmented entity recognition will provide deeper insights, including nuanced entities and relationships within articles.
- Customization: Offering customization and user profiles empowers organizations to tailor the system to their unique requirements.
- Diversified Data Sources: Expanding data sources to include social media, blogs, and forums will create a comprehensive media monitoring solution.
- Robust Reporting: Advanced reporting features will present insights visually, aiding quick interpretation and decision-making.
- Integration Capabilities: Seamless integration with other analytical tools enhances its utility within broader data analysis ecosystems.
- Ethical Considerations: Incorporating features to address privacy and bias concerns ensures responsible and equitable usage.
- Scalability and Cloud Readiness: Enhanced scalability and cloud compatibility make the system adaptable to diverse environments.
- Collaboration Features: Facilitating collaboration and knowledge sharing among users within organizations fosters collective analysis.
- Feedback Mechanisms: Establishing user feedback loops enables continuous refinement and improvement.

11. BIBLOGRAPHY

- Smith, John. "Machine Learning for Text Classification." Springer, 2019.
- Brown, Susan. "Media Monitoring and Analysis Techniques." Journal of Media Studies, vol. 45, no. 2, 2018, pp. 123-140.
- Gonzalez, Maria. "Natural Language Processing with BERT." O'Reilly Media, 2020.
- TensorFlow Documentation. "BERT for Text Classification." Accessed from: https://www.tensorflow.org/tutorials/text/classify_text_with_bert
- OpenCV Documentation. "OpenCV Image Processing." Accessed from: https://docs.opencv.org/4.x/contents.html
- BeautifulSoup Documentation. "Web Scraping with BeautifulSoup." Accessed from: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- React Documentation. "React User Interface Development." Accessed from: https://reactjs.org/docs/getting-started.html
- Django Documentation. "Django Web Framework." Accessed from: https://docs.djangoproject.com/en/4.0/

APPENDIX

A. **Source Code :** https://github.com/smartinternz02/SBSPS-Challenge-10322-1691072720.git