Automated Extraction of Key Issues from News

Name: Niranchna Natarajan
Register no. 22BCE5096
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, Tamil Nadu

Name: Keerthana Jayaprakashan
Register no. 22BCE5107
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, Tamil Nadu

Name: Esther Rachel Thomas
Register no. 22BCE5120
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, Tamil Nadu

Abstract — **In today's fast-paced information landscape, extracting meaningful and verified problem statements from news sources is vital for proactive decision-making and innovation. This paper presents a web-based solution for problem statement extraction from newspapers, aiming to support industry, government, and researchers in identifying pressing societal challenges. The platform enables users to upload newspaper PDFs, from which text is extracted and analyzed for authenticity using a fake news detection model. Verified news items are further processed by a text classification model to identify problem statements. These statements are then categorized into four domains—World, Sports, Business, and Science/Technology—using a specialized domain classifier. The result is an organized collection of real-world problem statements across key domains, providing actionable insights for companies, researchers, and policymakers to develop innovative solutions. By offering a continuous feed of current challenges, this tool aids in informed decision-making and fosters timely responses to global and regional issues. The website's user interface includes an upload area and category-specific buttons, allowing users to view categorized problem statements by domain, thus improving accessibility and application.**

Keywords: problem statement extraction, fake news detection, text classification, domain classification, societal challenges, real-world problems, decision support system, automated analysis

I. INTRODUCTION

With the rise in fake news and the growing challenge individuals face in identifying genuine problem statements from current news, there is an urgent need for automated systems that can filter credible information and highlight critical societal issues. In today's information-saturated environment, misinformation can spread rapidly, leading to confusion, misplaced priorities, and uninformed decision-making. This issue is particularly pressing when it comes to identifying real-world problems that require immediate attention, as the noise of sensational or misleading content often obscures genuine challenges that need solutions.

The project introduces an AI-based solution designed to address these issues by processing news articles in PDF format to assess their credibility and identify key problem statements. The system first extracts text from uploaded PDFs, which is then subjected to a fake news detection algorithm. This step filters out misleading content, ensuring that only reliable news articles proceed to further analysis. For those contents deemed credible, the system performs an additional check to identify whether they contain discussions of societal problems, such as environmental threats, economic instability, or technical risks and classifies them as problem statements.

Once a problem is identified, the system categorizes it based on its domain, helping users quickly locate issues within specific areas. This classification enables stakeholders to better understand the types of challenges facing society and allows for targeted responses to pressing concerns. By automating the process of verifying information and categorizing issues, this system supports data-driven decision-making, helping individuals and organizations focus on real-world problems that demand attention.

Overall, the project aims to enhance the quality and relevance of information sourced from news media by filtering out misinformation and bringing to light important societal issues. In doing so, it offers a valuable tool for researchers, policymakers, and organizations looking to address critical challenges based on accurate, reliable data. This system ultimately contributes to a more informed public and a proactive approach to solving complex global problems.

II. LITERATURE SURVEY

1. Jeelani Ahmed and Muqeem Ahmed (2021) [1], developed a text classification framework utilizing a Bayesian classifier, achieving a 93% accuracy rate, as evaluated through confusion metrics. Text classification is essential for tasks like text retrieval, summarization, and question-answering, where documents may be single-label or multi-label depending on their categorization. The classification process includes steps such as feature selection, document representation, algorithm application, and performance evaluation. Common algorithms used in this domain include Neural Networks, Support Vector Machines, k-Nearest Neighbour, and Naïve Bayes. The researchers trained and tested their model on a dataset containing approximately 75,000 news articles from the HuffPost and further validated it with articles from various live news websites. Among the algorithms tested, the Naïve Bayes classifier outperformed others, including KNN, which showed the lowest accuracy at 72%.

2. The paper published by the authors Manvendra Singh Chhajerh, Annanya KVS, Prof. Merin Meleet, Dr. Rajashekara Murthy S (2021) [2], focuses on the task of real-time news classification based on news headlines using machine learning techniques. The researchers aimed to develop an accurate model that can automatically classify news headlines into predefined categories. The paper evaluated and compared the performance of several machine learning algorithms including Multinomial Naive Bayes, Logistic Regression, Support Vector Machine, and Neural Networks. Metrics like precision, recall, and F1-score were

used to analyse the models' performance on different news categories. The results showed that Support Vector Machine and Logistic Regression achieved the highest accuracy, outperforming the other classifiers. The researchers then created a hybrid model combining the best-performing SVM and LR classifiers, which further improved the overall accuracy to 89.79%.

3. SzabóNagy and Kapusta (2023) [3] proposed a novel feature extraction technique, Tw-Idw (Term weight–inverse document weight), for fake news classification using natural language processing (NLP) methods. The study evaluates the effect of the Tw-Idw technique on the accuracy of text classification models by comparing it with the traditional Tf-Idf (Term frequency-inverse document frequency) method. Datasets related to COVID-19 and political news are analysed using machine learning models such as random forest and feedforward neural networks. The research mainly focuses on the importance of syntactic and morphological analysis in NLP tasks and highlights the ability of the Tw-Idw technique to enhance the performance of text classification models. The results demonstrate that the Tw-Idw technique has the potential to improve the efficiency of NLP-based text classification models, specifically in the context of detecting fake news, with an accuracy rate of up to 93.6%.

4. The paper published by the authors Amaram Divija, Kurkuri Smitha Kiran, Telukuntla Priyanka, Dr. Mantha Shailaja (2022) [4], resents a fake news detection model based on LSTM (Long Short-Term Memory) and Bi-LSTM (Bidirectional Long Short-Term Memory) deep learning techniques. The main objective was to use only the title content of news articles, without considering author features or other characteristics, to detect fake news. They pre-processed the data by tokenizing the text, converting to lowercase, removing punctuation, and removing stop words. Both the LSTM and Bi-LSTM models achieved similar high accuracy, around 93%, in classifying the news articles as fake or real based on the title alone. The researchers also developed a web interface that allows users to input a title and get the model's prediction on whether the news is fake or not.

5. Li et al. (2016) [5] introduce a novel web news extraction framework that uses text detection techniques to overcome the limitations of traditional content extraction methods. Previous methods, like manual parsing rules, and machine learning-based models, often face challenges with the diverse and dynamic nature of web page layouts. With the help of video text detection methods, the proposed system converts HTML code into a projection profile using compound text-tag difference (CTTD) statistics, effectively identifying and filtering relevant text blocks. Evaluations with a large, multilingual dataset show that this method is more efficient than the existing techniques with a higher accuracy rate of 46.38% compared to Boilerpipe's 21.54%, and it also processes data 16.91 times faster than NReadability. This innovative approach provides a solution to the challenges in constructing large-scale text documents, providing a more efficient and scalable solution for extracting news content from varied web sources.

III. METHADOLOGY

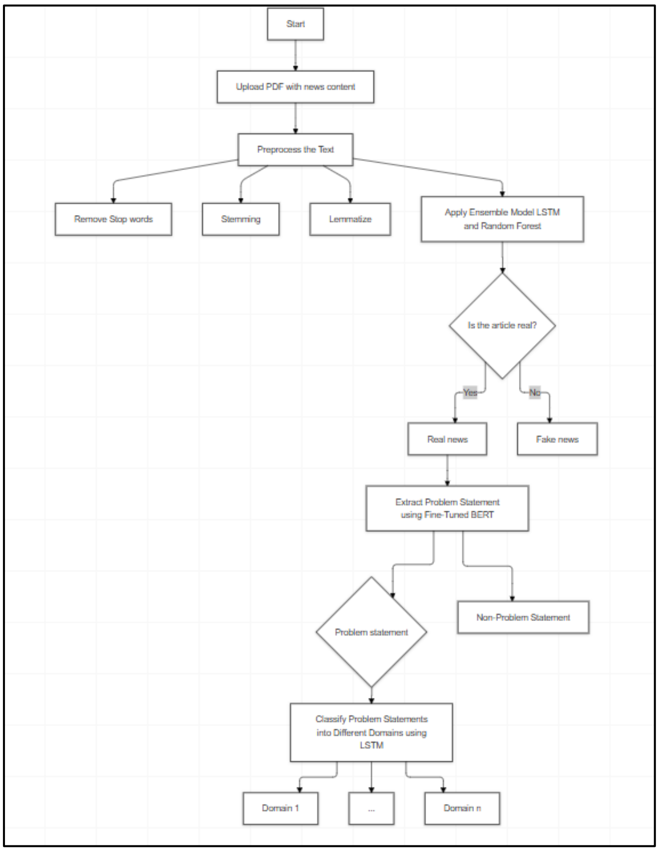
To address problem statement classification, domain classification, and real-fake news classification, we employed tailored deep learning methodologies for each task, focusing on efficient data preprocessing, feature extraction, model training, and deployment. Text data was cleaned by removing noise, such as punctuation and stop words, and structured into labeled datasets for classification. For problem statements, the dataset was organized into two categories: Problem Statement and Non-Problem Statement. Similarly, domain classification used datasets like AG_NEWS, while real-fake news classification relied on True.csv and Fake.csv. Texts were tokenized, with padding and truncation applied to standardize input lengths.

For problem statement classification, we fine-tuned a pre-trained BERT model for sequence classification. Texts were tokenized using the BERT tokenizer, generating token IDs, attention masks, and labels. A custom dataset class facilitated the management of these features during training. For domain classification, a multi-layered LSTM network was employed. It featured an embedding layer for semantic representations, three LSTM layers for capturing sequential dependencies, and a fully connected output layer for class probabilities. Similarly, for real-fake news classification, a Word2Vec-based embedding layer was combined with an LSTM model, followed by a Random Forest classifier to create an ensemble model for robust predictions.

All models were trained with cross-entropy loss and optimized using Adam. Training loops processed data in batches, leveraging GPUs for efficiency. Validation datasets were used to monitor performance, with metrics such as accuracy, precision, and F1-score guiding model refinement. The real-fake news classifier further benefited from ensemble learning, combining LSTM and Random Forest predictions through weighted averaging.

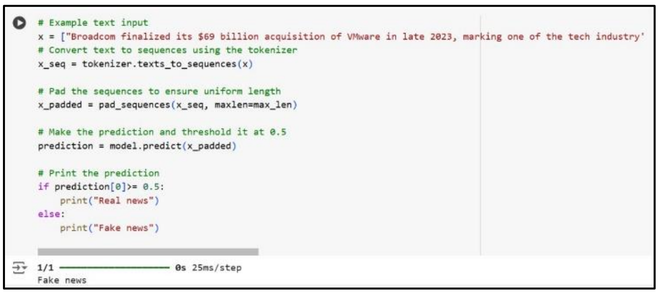
Post-training, models were saved with their respective tokenizers and preprocessing pipelines for real-time inference. By combining advanced NLP techniques, such as BERT for contextual understanding, LSTMs for sequential dependencies, and Word2Vec embeddings for semantic richness, these methodologies efficiently tackled classification tasks across diverse domains. The resulting models are optimized for deployment, ensuring reliable classification of unseen data.

IV. ARCHITECTURE


Start
Upload PDF with news content
Preprocess the Text
Remove Stop words
Stemming
Lemmatize
Apply Ensemble Model LSTM and Random Forest
Is the article real?
Yes
No
Real news
Fake news
Extract Problem Statement using Fine-Tuned BERT
Problem statement
Non-Problem Statement
Classify Problem Statements into Different Domains using LSTM
Domain 1
...
Domain n

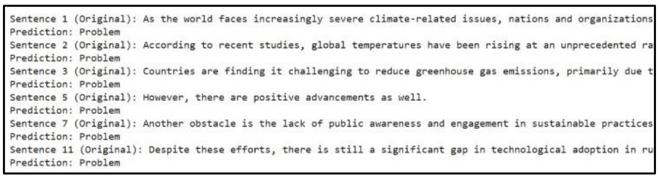
V. RESULTS

The system successfully processes news articles in PDF format by extracting and analysing the text to identify relevant information. Initially, the PDF file is ingested, and the textual content is extracted for further processing. The extracted text undergoes a fake news detection phase, where any identified fake news is filtered out, ensuring that only credible information proceeds to the next steps.



```
# Example text input
x = ["Broadcom finalized its $69 billion acquisition of VMware in late 2023, marking one of the tech industry'
# Convert text to sequences using the tokenizer
x_seq = tokenizer.texts_to_sequences(x)

# Pad the sequences to ensure uniform length
x_padded = pad_sequences(x_seq, maxlen=max_len)

# Make the prediction and threshold it at 0.5
prediction = model.predict(x_padded)

# Print the prediction
if prediction[0]>= 0.5:
    print("Real news")
else:
    print("Fake news")

1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 25ms/step
Fake news
```

The credible statements are then divided into problem and non-problem categories, allowing the system to focus on statements that highlight pressing societal issues.



```
Sentence 1 (Original): As the world faces increasingly severe climate-related issues, nations and organizations
Prediction: Problem
Sentence 2 (Original): According to recent studies, global temperatures have been rising at an unprecedented ra
Prediction: Problem
Sentence 3 (Original): Countries are finding it challenging to reduce greenhouse gas emissions, primarily due t
Prediction: Problem
Sentence 5 (Original): However, there are positive advancements as well.
Prediction: Problem
Sentence 7 (Original): Another obstacle is the lack of public awareness and engagement in sustainable practices
Prediction: Problem
Sentence 11 (Original): Despite these efforts, there is still a significant gap in technological adoption in ru
Prediction: Problem
```

Finally, the identified problem statements are categorized into distinct domains, such as environment, health, economy, or technology, facilitating a structured understanding of issues across different sectors. This streamlined process enhances the relevance and categorization of societal challenges, providing a valuable tool for informed decision-making.

VI. CONCLUSION

The proposed system offers a comprehensive solution for extracting and classifying problem statements from newspaper content through the integration of multiple advanced models. By leveraging fake news detection, text classification, and domain-specific categorization models, the platform ensures that only authentic news is processed and analyzed for real-world challenges. The process begins with a fake news detection model, which analyzes the uploaded PDF to ensure that only authentic, credible news articles are processed. This model is trained to identify patterns and inconsistencies in the text that are commonly associated with misinformation, such as sensationalism, bias, and unreliable sources. Once the news content is verified as authentic, it is passed through two text classification models for further processing.

The first classification model identifies whether the content contains a problem statement or not. If a problem statement is detected, the second model categorizes it into one of four predefined domains: World, Sports, Business, or Science/Technology. This domain classification model leverages natural language processing (NLP) techniques to analyze the context of the problem and assign it to the appropriate domain based on its content.

The results demonstrate the system's ability to efficiently and accurately classify problem statements, providing users with actionable insights that can aid in decision-making. By automating this process, the platform offers organizations, researchers, and governments a valuable tool for staying informed on current societal challenges and fostering solutions to address them.

VII. LIMITATIONS OF THE STUDY

While the proposed system offers significant potential for automating the extraction and classification of problem statements from newspaper content, several limitations must be considered. These challenges could impact the system's effectiveness, accuracy, and applicability in real-world scenarios. The key limitations include:

- *Accuracy of Fake News Detection:* Fake news detection algorithms may struggle with nuanced or satirical content, leading to potential false positives or negatives. High accuracy in distinguishing fake from real news requires sophisticated models, which can be challenging to maintain consistently.

- *Text Extraction Quality:* Extracting text accurately from diverse newspaper formats and PDF layouts, especially if the documents include images, complex formatting, or low resolution, can result in errors that affect subsequent analysis.

- *Classification Challenges*: The models used for identifying problem statements and classifying them into specific domains may face difficulties with ambiguous or overlapping topics. For instance, a problem related to technology in sports may be difficult to categorize accurately.

- *Contextual Limitations:* The system might lack the ability to recognize the context and underlying nuances of certain issues, which can lead to oversimplified classification or misinterpretation of the problem's scope.

- *Dependence on Model Training Data:* The quality and representativeness of the data used to train the classification models significantly impact performance. Biases in training data can lead to misclassification or an incomplete representation of real-world problems.

- *Dynamic Nature of News:* News is constantly evolving, and models need regular updates to stay relevant and adapt to new terminologies, trends, and issues. Without consistent updates, the system may struggle to classify novel or emerging issues.

- *Privacy and Data Security:* Handling news content, especially if it's sensitive, requires strong data security measures. Ensuring the privacy and integrity of the uploaded files and the processed information is crucial.

- *Scalability and Performance:* Processing large PDF files and running multiple models in sequence can be resource-intensive, potentially leading to performance issues, especially if there is high user demand.

VIII. FUTURE SCOPE

The development of the project lies in enhancing the accuracy and sophistication of both the fake news detection and problem identification mechanisms. With advances in natural language processing (NLP) and machine learning, future iterations of the system could incorporate more advanced algorithms for better sentiment and intent analysis, allowing it to better distinguish between sensationalism and factual reporting. Integrating real-time data sources beyond PDFs, such as news websites and social media feeds, using web scraping methods, could also help broaden the scope, enabling the system to identify emerging issues as they develop, thus enhancing its timeliness and relevance. Furthermore, the system could be improved to handle multilingual news sources, making it adaptable to diverse regions and able to capture a global perspective on pressing issues.

Another promising direction involves collaboration with other AI technologies, such as predictive analytics and trend forecasting, to identify not only current but also potential future societal challenges based on historical data and current patterns. By integrating with visualization tools, the system could provide interactive dashboards that allow users to explore trends and connections across various problem domains, making it a valuable asset for policymakers, researchers, and organizations aiming to anticipate and address problems proactively. Additionally, expanding the system to suggest actionable insights or solutions tailored to specific problem categories could transform it from an analytical tool into a strategic resource for driving meaningful societal impact

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our mentor, Dr. Sharmila Devi, for their invaluable guidance, encouragement, and support throughout this project. Their insights and expertise have been instrumental in shaping our research and have greatly contributed to the successful completion of this work. We are also thankful to our colleagues and team members for their collaborative efforts and constructive feedback, which enriched our project development process. Finally, we extend our appreciation to Vellore Institute of Technology for providing us with the resources and facilities needed to carry out this research. This project would not have been possible without the collective support and encouragement we received.

REFERENCES

[1] Ahmed, J., & Ahmed, M. (2021). ONLINE NEWS CLASSIFICATION USING MACHINE LEARNING TECHNIQUES. IIUM Engineering Journal, 22(2), 210–225. https://doi.org/10.31436/iiumej.v22i2.1662

[2] Chhajerh, M. S., KVS, A., Meleet, M., S, R. M., & Department of Information Science and Engineering, RV College of Engineering, Bengaluru. (2021). Real Time News Headlines Classification Using Machine Learning. International Research Journal of Engineering and Technology (IRJET), 3296. https://www.irjet.net

[3] Nagy, K. S., & Kapusta, J. (2023). TwIdw—A Novel Method for Feature Extraction from Unstructured Texts. Applied Sciences, 13(11), 6438. https://doi.org/10.3390/app13116438

[4] Divija, A. (2022). Fake news classifier. International Journal for Research in Applied Science and Engineering Technology, 10(6), 1716–1722. https://doi.org/10.222

[5] Wu, Y. C. (2016b). Language independent web news extraction system based on text detection framework. Information Sciences, 342, 132–149. https://doi.org/10.1016/j.ins.2015.12.025