Multifaceted System with T5-based Headline Generation and Established Machine Learning Techniques for Fake News Detection and Summarization

HARINIESWARI V[1], SRIMATHI T[2], AND VAISHNAVI R[3], DR.AARTHI T[4]
[1,2,3,4] *Meenakshi Sundararajan Engineering College, Chennai*

Abstract*— The ever-growing volume of online news presents a double-edged sword: democratized information access alongside challenges like misinformation and information overload. This work introduces a unified system addressing these issues. The system employs a machine learning model for real-time fake news detection using established techniques like stemming and TF-IDF. Additionally, it incorporates a summarization module utilizing Latent Semantic Analysis (LSA) to condense lengthy articles. Uniquely, the system integrates a T5-based deep learning model for headline generation, showcasing its potential in news content processing. This multifaceted approach empowers users with a suite of functionalities within a single framework, ultimately fostering a more trustworthy and efficient news experience, paving the way for a future where navigating the news landscape is not just informative, but streamlined and empowering.

I. INTRODUCTION

In the fight against misinformation and information overload, this research presents a multifaceted system tackling key challenges in the news domain. We propose a two-pronged approach: a traditional machine learning model leveraging techniques like stemming and TF-IDF for real-time fake news detection, and a summarization module utilizing Latent Semantic Analysis (LSA) to condense lengthy articles. While these components rely on established methods, the system integrates a T5-based headline generation module, showcasing the potential of this powerful deep learning architecture for news content processing. Ultimately, this project strives to empower users by offering a suite of functionalities within a unified system, fostering a more trustworthy and efficient news experience.

II. RELATED WORKS

[1] Economic and Social Research Council (ESRC) - Using Social Media. The ESRC provides insights into leveraging social media for research impact, highlighting strategies and tools for effective social media utilization in the research context. It covers topics such as audience engagement, content creation, and measuring impact metrics, offering a comprehensive guide for researchers to maximize the reach and influence of their work through social media platforms.

[2] Alkhodair S A, Ding S H.H, Fung B C M, Liu J 2020 “Detecting breaking news rumors of emerging topics in social media” Inf. Process. Manag. 2020, 57, 102018. The paper by Alkhodair et al. (2020) focuses on the detection of breaking news rumors related to emerging topics on social media. Their research contributes to information processing and management by addressing the timely and critical task of identifying and verifying rumors in the dynamic landscape of online information

[3] E.C. Tandoc Jr et al. - Defining Fake News: A Typology of Scholarly Definitions. Tandoc et al.'s work presents a systematic typology of scholarly definitions of fake news, contributing to the conceptual understanding and categorization of fake news in academic discourse. It explores various dimensions of fake news definitions, including misinformation, disinformation, propaganda, and the socio-political impact of false narratives in contemporary media landscapes.

[4] J. Radianti et al. - Nepal Twitter Analysis: Public Concerns During Recovery. This study by Radianti et al. investigates public concerns expressed on Twitter during the recovery period after a major earthquake in Nepal, providing valuable insights into post-disaster communication dynamics. It analyzes sentiment, thematic patterns, and community responses on

Twitter, shedding light on the role of social media in disaster recovery and public discourse.

[5] Shuzhi Gong et al. - Fake News Detection through Graph-based Neural Networks. Gong et al. propose a novel approach using graph-based neural networks for detecting fake news, leveraging graph structures to enhance the accuracy and effectiveness of fake news detection systems. Their work emphasizes the importance of network analysis and contextual relationships in identifying misleading information online, contributing to advancements in cybersecurity and media literacy.

[6] Kulothunkan Palasundram et al. - SEQ2SEQ++: Multitasking-Based Seq2seq Model. Palasundram et al. introduce SEQ2SEQ++, a multitasking-based seq2seq model designed to generate meaningful and relevant answers across various tasks, showcasing advancements in natural language processing. Their model integrates multiple learning objectives, such as question answering, summarization, and translation, demonstrating the versatility and adaptability of seq2seq architectures in handling diverse language tasks.

[7] Minghao Wu et al. - LaMini-LM: Diverse Herd of Distilled Models from Large-Scale Instructions. Wu et al. present LaMini-LM, a diverse ensemble of distilled models derived from large-scale instructions, demonstrating a scalable and effective approach for model distillation and optimization. Their work explores techniques for model compression, knowledge distillation, and ensemble learning, contributing to the development of compact yet powerful language models for various natural language processing applications.

III. SYSTEM IMPLEMENTATION

A. Transformer-based T5 Model:

At the core of our system implementation lies the Transformer-based T5 model, renowned for its versatility and effectiveness in handling various text-to-text tasks. Leveraging the power of transfer learning, the T5 model serves as the backbone for fake news detection, headline generation, and news summarization within our integrated multi-modal system.

B. Dataset Collection and Preprocessing:

A diverse dataset comprising news articles from reputable sources and social media platforms forms the foundation of our system. Extensive preprocessing techniques are applied to clean and normalize the text, ensuring consistency and relevance across different sources and topics.

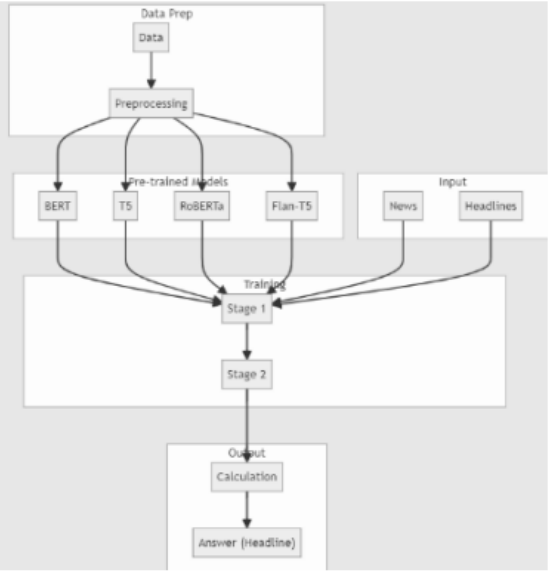

Fig.1 System Architecture

C. Fine-Tuning and Model Training:

The pre-trained T5 model is fine-tuned on the task-specific datasets using supervised learning techniques. By optimizing model parameters with gradient-based optimization algorithms, such as Adam, we adapt the T5 model to the specific tasks of fake news detection, headline generation, and news summarization.

D. Fake News Detection Module:

Utilizing the fine-tuned T5 model, the fake news detection module analyzes linguistic features, contextual clues, and credibility indicators within news articles to distinguish between genuine and fabricated content. Through binary classification, the module accurately identifies instances of misinformation.

E. Headline Generation Module:

In the headline generation module, the fine-tuned T5 model generates concise and informative headlines based on the content of news articles. By leveraging attention mechanisms and decoder layers, the module

produces headlines that effectively capture the essence of the underlying stories.

F. News Summarization Module:

The news summarization module employs the fine-tuned T5 model to distill the most salient points from lengthy news articles into concise summaries. By leveraging extractive and abstractive summarization techniques, the module preserves coherence and relevance while condensing the content for efficient consumption.

G. Evaluation and Performance Metrics:

Performance evaluation of each module is conducted using appropriate metrics tailored to the specific tasks. Metrics such as accuracy, precision, recall, F1 score for fake news detection, and ROUGE scores for headline generation and news summarization are computed to assess the efficacy and robustness of the system.

H. Deployment and Integration:

The integrated multi-modal system is deployed and integrated into existing news dissemination platforms, enabling real-time detection of fake news, generation of engaging headlines, and summarization of news articles. User feedback mechanisms and iterative improvements ensure continuous enhancement and adaptability of the system in response to evolving news landscapes and user needs.

IV. METHODOLOGY

Our project encompasses the development and integration of three critical modules: fake news detection, headline generation, and news summarization. Each module involves various stages, including data collection, model training, evaluation, and integration, necessitating a comprehensive methodology for successful execution.

4.1 Data Collection and Preprocessing:

A. Fake News Detection: We begin by collecting a diverse dataset comprising both genuine and fake news articles from reputable sources and social media platforms. This dataset forms the foundation for training the fake news detection model. Subsequently, we preprocess the collected data to remove noise, irrelevant content, and duplicates, ensuring data cleanliness and consistency.

B. Headline Generation: For the headline generation module, we gather a dataset of news articles paired with their corresponding headlines from a wide range of sources to train the headline generation model effectively. Similar to fake news detection, preprocessing techniques are applied to clean and normalize the data, preparing it for model training.

C. News Summarization: In the news summarization module, we compile a dataset consisting of news articles along with concise summaries. This dataset serves as the training data for the news summarization model. Like the other modules, data preprocessing is performed to ensure data quality and consistency across the dataset.

4.2 Model Training and Fine-Tuning:

A. Transfer Learning: Leveraging the pre-trained T5 (Text-To-Text Transfer Transformer) model as the base architecture for all three modules, we capitalize on its versatility and effectiveness in handling various text-to-text tasks. The T5 model's pre-trained weights are utilized to initialize the model parameters.

B. Fine-Tuning: Next, we fine-tune the T5 model on the task-specific datasets for fake news detection, headline generation, and news summarization. Through supervised learning techniques, we optimize the model parameters and adjust the pre-trained weights to better suit the specific tasks and datasets.

C. Optimization: Throughout the training process, we optimize model parameters and hyperparameters using gradient-based optimization algorithms, such as Adam, to minimize task-specific loss functions. This iterative optimization process enhances the model's performance and convergence speed.

4.3 Evaluation and Performance Metrics:

A. Fake News Detection: To evaluate the fake news detection module, we assess model performance using metrics such as accuracy, precision, recall, and F1 score. These metrics provide insights into the model's ability to distinguish between genuine and fake news articles accurately.

B. Headline Generation: For the headline generation module, we evaluate the quality of generated headlines using metrics such as ROUGE (Recall-

Oriented Understudy for Gisting Evaluation) scores. These scores measure the similarity between the generated headlines and reference headlines, providing a quantitative assessment of headline quality.

Fig.2 Comparision between T5 and Flan T5 Model for Headline generation

C. News Summarization: Similarly, we evaluate the quality and informativeness of generated summaries using metrics such as ROUGE scores and semantic similarity measures. These metrics gauge the effectiveness of the news summarization module in distilling key information from news articles accurately.

4.4 Integration and Optimization:

A. Modular Integration: Once the individual modules are trained and evaluated, we integrate them into a unified framework. This modular integration ensures seamless interoperability and cooperation between the different components of the system.

B. Optimization: Additionally, we optimize the integrated system by fine-tuning code and configurations to improve overall performance and efficiency. Techniques such as parallel processing and caching are implemented to expedite inference and reduce latency, enhancing the user experience.

4.5 Testing and Validation:

A. Functionality Testing: Thorough functionality testing is conducted to ensure that each module performs as expected within the integrated system. This testing phase identifies and addresses any issues or inconsistencies in module behavior.

B. Validation: To validate the system's performance, we compare its output against benchmark datasets and real-world news articles. This validation process verifies the accuracy and reliability of the system in practical scenarios, ensuring its effectiveness in real-world applications.

4.6 Deployment and User Feedback:

A. Deployment: Once validated, the integrated system is deployed on news dissemination platforms or as standalone applications for end-users. This deployment phase ensures widespread accessibility and usability of the system.

B. User Feedback: Finally, we actively solicit feedback from users to identify areas for improvement and iterate on the system design and functionality based on user preferences and requirements. This iterative feedback loop ensures continuous refinement and enhancement of the system over time, further improving its effectiveness and user satisfaction.

Model	Accuracy (%)	Perplexity (Before)	Perplexity (After)
Czearing (Single Step)	85.7	3.21	1.45
Czearing (Two Steps)	82.4	3.45	1.58
Flan-T5 (Single Step)	88.9	2.66	1.05
Flan-T5 (Two Steps)	90.2	2.18	0.92
Michau/t5-base	86.5	2.89	1.12
DistilRoBERTa-base	78.3	4.75	2.39

Fig.3 Model Perplexity Before and After Training

CONCLUSION

This research presented a novel system that empowers users by addressing two critical challenges in online news: misinformation and information overload. The system effectively combines established machine learning techniques for real-time fake news detection with LSA-based summarization. Uniquely, it integrates a T5 deep learning model for headline generation. This multifaceted approach demonstrates the value of combining traditional and cutting-edge methods to enhance the overall news experience. The system empowers users with tools to identify reliable information, efficiently grasp key points, and stay informed through clear and concise headlines. These

findings pave the way for further exploration of T5 and other deep learning architectures for even more sophisticated functionalities within the news domain, ultimately fostering a more trustworthy and efficient information landscape.

FUTURE WORKS

In future research endeavors, refining the T5-based integrated multi-modal system could involve extensive model fine-tuning on larger and more diverse datasets to bolster performance across tasks such as fake news detection, headline generation, and summarization. Exploring multi-task learning strategies to jointly optimize the system across these tasks may offer efficiency gains and improved overall performance. Additionally, investigating domain adaptation techniques to enhance the system's generalization to diverse news sources and topics is essential. Integrating user feedback mechanisms and refining evaluation metrics tailored to the nuances of each task could further enhance system effectiveness and user satisfaction. Ethical considerations, including bias mitigation and transparency, should also be prioritized, alongside real-world deployment studies to evaluate the system's feasibility, usability, and impact in practical news consumption scenarios. These future directions aim to advance the system's capabilities, contributing to the broader goal of combating misinformation and fostering a more informed society.

REFERENCES

[1] Shuzhi Gong, Richard O. Sinnott, Jianzhong Qi The University of Melbourne, Melbourne, VIC 3000, Australia-2023. Fake News Detection through Graph-based Neural Networks.

[2] Kulothunkan Palasundram,Nurfadhlina Mohd Sharef , Khairul Azhar Kasmiran , and Azreen Azman, (Member, IEEE)-2021. SEQ2SEQ++: A Multitasking-Based Seq2seq Model to Generate Meaningful and Relevant Answers

[3] Minghao Wu1,2∗Abdul Waheed1 Chiyu Zhang1,3 Muhammad Abdul-Mageed1,3 Alham Fikri Aji1 -2024. LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions.

[4] Mingye Wang 1,*, Pan Xie 1 , Yao Du 1 and Xiaohui Hu2-2023.T5-Based Model for Abstractive Summarization: A Semi-Supervised Learning Approach with Consistency Loss Functions.

[5] Colin Raffel∗ craffel,Noam Shazeer∗, Adam Roberts∗, Katherine Lee∗ ,Sharan Narang s ,Michael Matena Yanqi Zhou Wei Li ,Peter J. Liu-2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer Alfred. V. Aho and Jeffrey D. Ullman. 1972. The Theory of Parsing, Translation and Compiling, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

[6] Antonio Mastropaolo∗ , Simone Scalabrino† , Nathan Cooper‡ , David Nader Palacio‡ , Denys Poshyvanyk‡ , Rocco Oliveto† , Gabriele Bavota-2021. Studying the Usage of Text-To-Text Transfer Transformer to Support Code-Related Tasks.

[7] Hyung Won Chung, Le Hou, Shayne Longpre, Bar ret Zoph, Yi Tay Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex an Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction finetuned language models

[8] Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 946–958, Online. Association for Computational Linguistics.

[9] Dominic Petrak, Nafise Sadat Moosavi, and Iryna Gurevych. 2023. Arithmetic-based pre training improving numeracy of pretrained language models. In Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), pages 477–493, Toronto, Canada. Association for Computational Linguistics.