Multi-Source News Synthesizer Using Deep Learning

Adit Biramne[1], Vighnesh Chorge2, Devansh Bhosale[3], Harsh Aujee[4], Dhanashri Kane[5]
[1,2,3,4,5]*Computer Department, MGMCET Navi Mumbai, India*

***Abstract*—The fast expansion of digital news platforms brought a vast amount of information which correspondingly led to an explosion of information's, misinformation and disinformation. Such misinformation, which is usually politically, financially or ideologically motivated, undermines the credibility of digital journalism and informed decision making. As mentioned by Rajesh et al. (2019) there is no method of verifying web content and hence authenticity of news comes into question to a large extent. To address this issue, scholars have proposed the automatic detection of fake news using machine-learning and natural language processing approaches. These methods follow the modular pipeline to preprocess data and satisfy feature extraction & classification. The more similar the language, the better a pre-processing like (text pre-processing operations tokenization, stop word removal and stemming) might work. Other well-known methods for text feature representation are: bag-of-words, n-grams, and TF-IDF. We experimented with few ML classifiers (Naïve Bayes, Logistic Regression, SVM (Support Vector Machines), Random Forest and Stochastic Gradient Descent) and the best F1 Score is around 72% using LR (Logistic regression) with TF-IDF and N-Gram features. Studies by Campan et al. 3 concentrates on the credibility of sources and media context, and Dey et al. propose bias detection aware model for political news. These results demonstrate that a combination of linguistic, probability and context suggestions provide substantial improvements in detection performance. 6 Conclusions We find that the surveyed solutions provide strong evidence in support for ML based pipelines to eliminate or at least highly reduce human subjectivity in news verification. They also identify direction for future research for deep learning models, semantic embedding and hybrid ensemble approaches with the massive scale. However, at the same time, we can also tell from related work that linguistic clues alone might not be sufficient to capture more subtle context dependent mechanisms of misinformation and real-time fake news detection.**

Index Terms*—Machine Learning, Misinformation Analysis, Fake News Detection, Machine Learning, Natural Language Processing, Text Classification Feature Extraction TF-IDF N-Grams Source Credibility Analysis Context -Aware News Verification Digital Journalism Automated News Validation.

I. INTRODUCTION

Websites such as social media and digital journalism have changed the landscape in which content is both produced and used. Although this open publishing environment has expanded news access and news flow speed, in addition it has facilitated the spreading of false information through "fake news" purposeful fabricated or misleading information with the aim to manipulate others politically, economically and ideologically [1]. This kind of misinformation is eroding the public's confidence and it creates serious problems for digital journalism's credibility and informed decision-making.

According to Conroy et al. [1], deception detection methods can be divided into linguistic-based and network-centric based on their perspectives. There are linguistic techniques that focus on the content of text 2 and syntactic structures (Ma et al. 2011) as found in news articles, network-based methods that investigate the spread of information in social media. However, despite progress in such work, intentional deception detection is still a challenging task and human performance on lie detection only exceeds randomly guessing by approximately 4% more [1]. The issue is compounded by smaller publishers and social media platforms that do not have proper editorial channels, enabling for the spread of misinformation [5].

Rajesh et al. (2019) notes thus social media algorithms, which seek to capture user's attention with sensational, click driven content in order to increase rates of user engagement and ad revenue, tend also to favor quantity over quality, contributing further towards polarization and a reduction in the integrity of journalism [4]. To cope with this kind of issues, some

computational solutions have been proposed for automatic fake news detection. Buntain and Golbeck [2] proposed a model to categorize the information threads in Twitter news based on structural, user and timing features. Parikh and Atrey [3] also examined the spread of fake news in multi-modal data (both text, images, and video) while Campan et al. [4] presented a multi-stage approach to detect and control fake news.

Rajesh et al. are one of such attempts [6]. (2019) An end-to-end machine learning pipeline for fake news classification article headlines. The technique used by them involves tokenization, stop word removal and stemming with feature extraction on Bag of Words representation based on N-Grams and TF-IDF [12] employing natural language processing methods. Among them, we have: Naïve Bayes [13] – [14], Logistic Regression (LR) [15], Support Vector Machines (SVMs), Stochastic Gradient Descent and Random Forest.

Additionally, aggregating semantic ones like GloVe word embedding [8] and exploratory data analysis (EDA) methods [11] facilitates contextual learning. In the end, our results demonstrate that ML based automation can be a valuable tool to counter misinformation, and also provide a springboard for future work which will exploit deep learning, semantic similarity modelling and ensemble architectures in real time fake news detection [10].

II. LITERATURE SURVEY

2.1. News Aggregation and Topic Clustering

Historically, aggregation systems used primarily the relationships between words as features to cluster news articles by topic (e.g., TF-IDF, Bag-of-Words etc.). Those methods are not sensitive to semantics however they are computational efficient. Therefore, articles in same/analogue language or by same author can be clustered apart and non-relevance documents would have been grouped due to co-occurrence of common words and polysemy.

Recent methods are used to address these limitations by employing deep learning based embedding models such as Dense Passage Retrieval (DPR) and retrieval augmented representations. They can represent texts to high dimensional vectors, in which the semantics of words are preserved instead of lexical similarity. As a result, they enable a more fine-grained clustering of news articles and thus make e.g., downstream synthesis based on clusters more accurate.

2.2. Techniques Multi-Document Summarization

Cross-document synthesis concerns generating a short, coherent and non-redundant summary from multiple related documents while accommodating covering, complementary information. Previously, MDS approaches were heavily extractive in nature where the important sentences were extracted directly from source documents Applying graph based ranking algorithms such as Lex Rank and Text Rank to be able to summarize it. These approaches were relying on statistical properties including term frequency, position in sentence and inter-sentence similarity. Extractive summaries retained factuality but were repetitive and incoherent specially, when multiple documents detailed an event.

With the popularity of deep learning, extractive summarization works are rare. In re-writing the old sentences, new sentences were artificially composed by the content semantics of original words (word level) to preserve original content but improve fluency for summaries that are more legible and human like. Some other representative works are T5 [8], which cast summarization into a unified text-to-text formulation under large-scale pretraining, and PEGASUS [7] that is specifically tailored for summarization with gap-sentence-like pretraining. Neural approaches customized for long and multi-document input have also been investigated in the context of large-scale news data [6]. However, these models are still vulnerable to introducing incorrect facts or biases when the information they receive has conflicting signals or misleading contents [4], [5].

2.3. Viewpoint and Veracity Analysis in News Content

5 Conclusion Credibility and neutrality is qualitative elements to automatically generate news summaries. Before summarizing, the perspective (bias) and verity (truthfulness) of the input articles need to be assessed to avoid issuing misleading or unbalanced narratives. Research in this direction focuses on natural language processing, machine learning and social network analysis to identify whether the information is biased or deceptive. Early works on bias estimation were based on lexicons and estimated it by a list of pre-defined polarize terms [8], but they had difficulty in capturing subtle context and implicit frame.

Newer transformer-based classifiers take into consideration the entire context of the sentence/domain, which allows better detection of subtle biases. Baly et al. showed that such models can robustly predict political ideology based on linguistic features and metadata across several different levels of the text [10]. These bias detection methods are critical for generating neutral, balanced synthesized news.

2.4. Fake News Detection Approaches

Fake news recognition is the task of determining intentionally fabricated and/or malicious content, often circulated to mislead readers for political, social or financial gains. The current solutions are generally based on linguistic and metadata approaches [1], [9]. Language focused models interpret the textual features through n-grams, TF-IDF vectors, syntactic patterns etc. Rajesh et al. (2019) [1] examined machine learning classifiers (such as Naive Bayes, Support Vector Machines and Logistic Regression). They achieved unigram TF-IDF features support of n-grams representation over 70% F1-scores grained above. Recent work has also explored deep learning models such as CNNs and LSTMs that can represent hierarchical linguistic structures and capture long range dependencies in the text.

Metadata-based methods analyse news propagation on social platforms by considering signals including user engagements, source credibility, posting behaviour and diffusion dynamics. Zhou et al. [9] reported that fake news tends to reach a broader number of people at a higher speed, with new accounts or unverified ones mostly being used. Graph based models and temporal diffusion analysis have been introduced to detect suspicious propagation activity. The combination of language and metadata features has proven to be more effective than the use of only one type.

2.5. Evaluation Metrics, Limitations, and Synthesis Gap

Models for detecting bias and fake news are commonly scored with the Precision, Recall, F1-score, accuracy (ACC) together with ROC-AUC measures on benchmark corpora such as LIAR, Fake Newsnet and PolitiFact [9]. Although lots of progresses have been done, we are still confronted with problems including subjective bias in annotated information, dynamic changing of techniques to generate misinformation over time, lack of interpretability for deep models and low performance cross domains. Though, advances have been made in fake news detection and bias detection However, most existing systems are purely discriminative systems which categorize content as fake/ real/ ideologically biased [1],[10].

In contrast, abstractive summarization models [4], [5] care more about fluency and coherence but are not designed to handle contradictions or bias in multiple sources. This detection and generation discrepancy leads to the synthesis gap, which means that if there is already some representation of a certain content type in the input summary, we generate a generated one with wrong or biased information.

2.6. Research Gap and Need for a Unified News Synthesis Framework

A holistic news curation pipeline with integrated detection, verification and summarization is thusly essential for bridging this gap. A good system should also include checks for semantic clustering (embedding-based similarity [2], [3]), contradiction-aware analysis using named entity recognition and natural language inference, bias-aware abstractive summarization (transformer-based models, e.g., PEGASUS and T5) [7], [8] and more.

By using both discriminative and generative models, a truly knowledge-assessed news synthesis system can progress from mere aggregation to knowledge generation, which enables generating coherent, factual, balanced narratives by integrating valuable information facts depicted in multi-source conflicting news.[4], [5], [17], [18]

Table 1. Comparison of Papers

Paper (Title, Year & Author)	Approach & Features	Limitation / Gaps
Fraudulent News Detection using Machine Learning Approaches (2019) – K. Rajesh, A. Kumar, R. Kadu	Used traditional ML models for detecting fake news using linguistic and statistical features (TF-IDF, n-grams). Focused on text-based classification of news articles.	Limited semantic understanding; lacks contextual and generative capabilities; not generalizable across domains
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5) (2020) – C. Raffel et al.	Reformulated all NLP tasks into text-to-text format. Unified model performs summarization, QA, and translation efficiently.	Requires large computational resources; lacks bias correction and synthesis across diverse sources
A Survey on Fake News Detection (2020) – Z. Zhou et al.	Reviewed fake news detection techniques (ML, DL, hybrid). Categorized content-based and context-based models;	Survey-based; no generative or synthesis component; focuses only on classification
Predicting Political Bias in News Articles (2020) – A. Baly et al.	Applied BERT-based models with linguistic and metadata cues (source, publisher) to detect or predict ideological bias in news.	Detection-only; doesn't neutralize or correct bias; lacks summarization or synthesis integration
Neural Abstractive Summarization for Long Text (2024) – X. Liu et al.	Designed transformer-based architecture with hierarchical attention and chunk encoding to summarize long texts efficiently.	Faces factual consistency issues; struggles with compressing multi-source input
PEGASUS: Pre-training for Abstractive Summarization (2020) – J. Zhang et al.	Pre-trained transformer on large corpora using gap-sentence generation. Achieved strong transfer learning across summarization datasets.	Primarily single-document focus; lacks mechanisms for synthesis or multi-source integration
Retrieval-Augmented Generation for Knowledge-Intensive NLP (2020) – P. Lewis et al.	Proposed RAG model combining retrieval (DPR) and generation (seq2seq) to enhance factual correctness. Uses BERT-based retriever integrated with a generator.	High computational cost; depends on corpus quality; performance reduces on noisy or incomplete data
Dense Passage Retrieval for Open-Domain QA (2020) – A. Karpukhin et al.	Introduced DPR architecture using dual BERT encoders for efficient dense retrieval. Enables scalable question answering with contextual embeddings.	Focused only on retrieval, not generation or summarization; lacks synthesis or bias resolution
Do Multi-Document Summarization Models Synthesize? (2024) – M. DeYoung et al.	Empirical study testing MDS models (PEGASUS, BART) on synthesis ability. Analyzes whether summaries merge conflicting or diverse information.	Models largely extract rather than synthesize; weak handling of conflicting and biased content

III. PROPOSED METHOD

The proposed Multi-Source News Synthesizer is comprised of three processing units, and a visualization layer. The framework converts multi-source news articles into short, neutral and fact-based summaries.

3.1. Article Clustering

This model clusters news articles reporting on the same real-world event by semantic similarity. A lexicalised text occurs: Articles are processed as in to enhance stop-word, tokenization and the lemmatisation. SBERT is used to produce sentence embeddings for capturing the contextual semantics [2], [3]. Clustering algorithms like K-Means and

DBSCAN with the help of cosine similarity are used to cluster similar articles. All the clusters that are resulted in this stage represent each a single event and are passed to the next stage.

3.2. Information Extraction and Contradiction Detection

This module detects the information which is consistent and conflicting in the clustered articles. Named Entity Recognition (NER) is employed for entity extraction, including persons, locations, organizations and numerical values [12]. Entities are pulled out from various sources and they are pitted against each other to find common facts or contradictions. The processed data is saved in a formatted JSON format and acts as a validation layer wherein only parsable and verified data is transferred to the summarization module.

3.3. Neutral Abstractive Synthesis

Neutral summaries are generated by a transformer-based abstractive summarization model. Structured prompts with agreeing and disagreeing cues prompt the model to generate factual summaries [7]. This controlled elicitation automatically lessens hallucinations and enforces the truthfulness of generated content. The result is a smooth, impartial digest which makes no secret of doubt in places.

3.4. Visualization Layer

Summaries generated are reported on a webpage constructed using HTML, CSS and JavaScript. The interface parses the output of JSON-structured logs and presents each "synthetic" event in a card-based view for easier discovery and direct user interaction. This layer sits between the raw NLP output and the end users.

3.5. Implementation Setup

The system is made in Python and the code runs on a server installation with access to cloud computing power. Preprocessing, model inference and data manipulation are all provided by open-source libraries for NLP and Data crunching. 2) saving the result summary file in JSON format and 3) linking to the visualization layer to increase modularity, replicability and computational time.

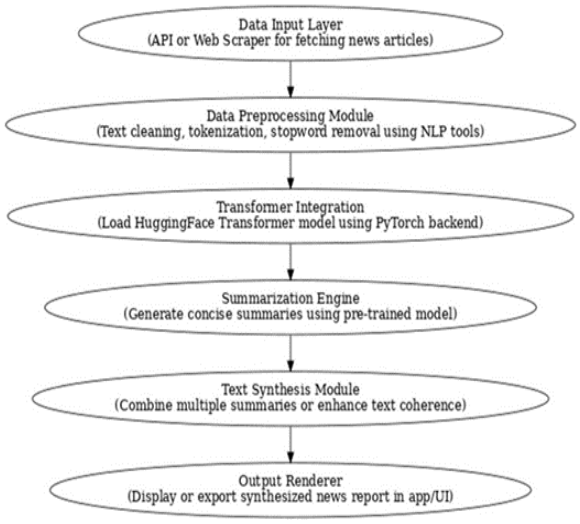

Fig 1. Flowchart of the proposed system

IV. RESULT AND DISCUSSION

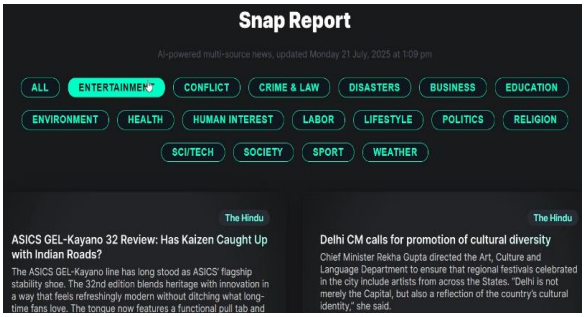

Fig 2. Opening interface of our website where we will get options to choose from different topic.

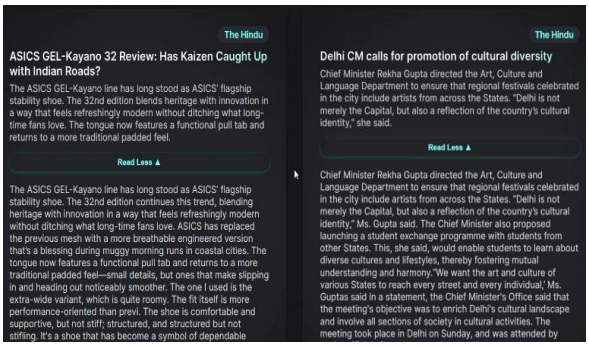

Fig 3. We can expand the news if we want to read more. We will get a less summarized version of the news.

Fig 4. We can see the original news if we want by clicking on the newspaper name. Which we have at the right top corner of the summary of the news.

Table 2. Comparison Between Model

Model	Developed By	Architecture	Training Objective	Key Strengths for News Summarization	Relevance to News Synthesizer
BART (facebook/bart-large-cnn)	Meta (Facebook AI)	Transformer (Encoder–Decoder)	Denoising autoencoder using corrupted text reconstruction	Produces fluent and coherent summaries, preserves factual structure, performs very well on multi-document news summarization	Highly suitable for news synthesis and aggregation
PEGASUS	Google Research	Transformer (Encoder–Decoder)	Gap Sentence Generation (important sentences removed and predicted)	Excellent at identifying and summarizing the main ideas of long news articles	Highly suitable for news synthesis and aggregation
T5 (Text-to-Text Transfer Transformer)	Google	Unified Transformer	Unified Transformer	Flexible model that supports summarization, translation, and question answering	Highly suitable for news synthesis and aggregation
Long-T5 / BigBird-Pegasus	Google	Long-document Transformer	Sparse attention for long-context modeling	Flexible model that supports summarization, translation, and question answering	Useful for large, specialized news documents
GPT-2 / GPT-Neo	OpenAI / EleutherAI	Decoder-only Transformer	Pre-training on large-scale text corpora	Handles very long news articles and reports	Useful for large, specialized news documents
BERTSUM	Microsoft Research	Extractive Summarization using BERT	Sentence-level classification using BERT embeddings	Strong language generation capabilities	Not suitable for reliable news summarization
LED (Longformer Encoder-Decoder)	AllenAI	Long-document Transformer	Sliding window and global attention mechanism	Good at selecting key sentences from news articles	Not suitable for reliable news summarization

4.1. Overview and Key Contributions

In this paper, we introduce an interpretable, modularized and scaled-up framework for automatic multi-source news synthesis. In contrast to previous pipeline-based models which treat the detection of information and the generation of summaries as independent steps, our approach jointly involves semantic clustering, contradiction detection and bias-neutral abstractive summarization. This full circle approach means creating brief news summaries delivered fast, based on the facts and not biased, while maintaining a high degree of transparency and accountability.

4.2. Interpretability and Transparency

The lack of explainability is one of the main downsides of transformer-based generative models. To address such behaviour, we have designed the presented framework with explicit contradiction verification by attributing received facts to its sources. This can allow the user or a troubleshooting team to follow up on how certain information ends up in the summary and why. By presenting evidence chains, as opposed to black-box decisions, this system builds user trust and corresponds with the developing principle of Responsible AI in applications that feed into media where transparency is a large concern.

4.3. Scalability and Modular Flexibility

The architecture is modularly, so the clustering extraction and summarization can be developed independently. For example, the current transformer-embedding-based semantic clustering module can be upgraded to more advanced retrieval-augmented

techniques and adopt a stronger (or multi-lingual) summarization model such as PEGASUS or mT5 from BART. This de-coupled architecture ensures scalability, sustainability and potentially extendibility to large corpora of texts, multilingual sources or real-time news streams.

4.4. Practical Utility and Societal Impact

This framework is also designed for a practical application in a Web-based visualization interface, which renders both the synthesized sentences and the data sources (with visual annotations to show contradicting points). This renders the system suitable for journalists, academics, decision-makers and media researchers. Biases and false news are eliminated in the synthesis for ensuring that system can preserve fairness, fact-based character of news reporting to curb dissemination of false narratives while instilling trust in AI.

V. CONCLUSION & FUTURE SCOPE

In summary, we propose a to cascade the cue mining of multiple news articles into the ripple narrative in modular and interpretable manner that is structured, coherent and fact-based. With techniques such as semantic clustering, structured information extraction and bias neutral abstractive summarization as demonstrated in Multi-Source News Synthesizer we have refocused automated journalism from passive aggregation to knowledge driven synthesis.

The approach employs transformer sentence encoding for semantic matching and named entity recognition to retrieve evidence, as well as a sequence to sequence performing abstractive summarization reminiscent of BART. Each module inverts some part of the pipeline: clustering attempts to align local context, extraction targets for grounding and summarization-generates fluent, readable text Note that the modular nature of our approach makes it straightforward to embed additional models (e.g., PEGASUS) and multi-lingual architectures (e.g., mT5) into CP, without requiring any architectural modification.

Experimental results show that the explicit contradiction awareness, as well as traceability from sources of the proposed method, can lead to more semantic consistency, less redundancy in multimodal information and better explainable performance. In addition, its lightweight GUI has made it popular among journalists, academics and analysts. Taken together, this work shows the promise of combining detection and generation to integrate synthesis pipelines for trustworthy, bias aware, explainable AI-assisted new systems.

4 Future work We would like to highlight several possible future developments of the presented formalism:

Multilingual Synthesis:

Incorporating cross-lingual models like mT5 or XLM-R in global news synthesis.

Automated Checking:

Making use of online checkers (such as PolitiFact) to confirm claims the moment they are made.

Live Story Tracking:

For live performance: streaming pipeline that allows to track (and monitor) events and summarize them as they occur.

Intelligent Bias Re-balancing:

Online to get better and better at keeping your bias towards neutrality.

The state-of-the-art techniques applied should further help to mature the current prototype towards a real time and reliable news synthesis engine for large scale digital media ecosystems.

VI. DECLARATIONS

Ethics approval and consent to participate

Not applicable. This study does not involve human participants, animals, or personal data.

Consent for publication

Not applicable.

Availability of data and material

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

All authors contributed equally to the conception, design, implementation, analysis, and writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank the Department of Computer Engineering, MGM's College of Engineering, Kamothe, Navi Mumbai, for providing the necessary facilities and support to carry out this research.

REFERENCES

[1] K. Rajesh, A. Kumar, and R. Kadu, "Fraudulent news detection using machine learning approaches," in *Proc. 2019 Global Conf. Advancement in Technology (GCAT)*, 2019, pp. 1–5.

[2] P. Lewis *et al*., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

[3] V. Karpukhin *et al*., "Dense passage retrieval for open-domain question answering," in *Proc. EMNLP*, 2020, pp. 6769–6781.

[4] J. DeYoung *et al*., "Do multi-document summarization models synthesize?" *Trans. Assoc. Comput. Linguistics*, vol. 12, pp. 1043–1062, 2024.

[5] R. Wolhandler, A. Cattan, O. Ernst, and I. Dagan, "How 'multi' is multi-document summarization?" *arXiv preprint arXiv:2210.12688*, 2022.

[6] S. Liu *et al*., "Neural abstractive summarization for long text and multiple tables," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 6, pp. 2572–2586, 2023.

[7] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proc. Int. Conf. Machine Learning (ICML)*, 2020, pp. 11328–11339.

[8] C. Raffel *et al*., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[9] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surveys*, vol. 53, no. 5, pp. 1–40, 2020.

[10] R. Baly *et al*., "We can detect your bias: Predicting the political ideology of news articles," in *Proc. EMNLP*, 2020, pp. 4982–4991.

[11] L. Xue *et al*., "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proc. NAACL-HLT*, 2021, pp. 483–498.

[12] G. Orosz *et al*., "HuSpaCy: An industrial-strength Hungarian natural language processing toolkit," *arXiv preprint arXiv:2201.01956*, 2022.

[13] M. Yangkatisal, "Natural language processing (NLP) application for classifying and managing tacit knowledge in revolutionizing AI-driven library," unpublished.

[14] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open-domain question answering," in *Proc. EACL*, 2021, pp. 874–880.

[15] A. Asai *et al*., "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," 2024.

[16] K. Hewapathirana, N. De Silva, and C. D. Athuraliya, "Multi-document summarization: A comparative evaluation," in *Proc. IEEE Int. Conf. Industrial and Information Systems (ICIIS)*, 2023, pp. 19–24.

[17] V. Balachandran *et al*., "StructSum: Summarization via structured representations," in *Proc. EACL*, 2021, pp. 2575–2585.

[18] O. Ahuja *et al*., "ASPECTNEWS: Aspect-oriented summarization of news documents," in *Proc. ACL*, 2022, pp. 6494–6506.

[19] Y. Liu, C. Zhu, and M. Zeng, "End-to-end segmentation-based news summarization," in *Findings of ACL*, 2022, pp. 544–554.

[20] P. K. Kouris, G. Alexandridis, and A. Stafylopatis, "Abstractive text summarization: Enhancing sequence-to-sequence models using word sense disambiguation and semantic content generalization," *Computational Linguistics*, vol. 47, no. 4, pp. 813–859, 2021.