Context-Aware Document Summarization System with Risk and Fake Data Detection

Mrs. K. Bhavadharani[1], Dr. P. Anbumani[2], Ms. S. Dhanavarshini[3], Ms. R. Divyadharshini[4], Ms. S. Dhoolika[5], Ms. C. Dharani[6]

[1]Asst. Prof/ Department of CSE, V.S.B Engineering College, Karur, Tamil Nadu
[2]Asst. Prof/ Department of CSE, V.S.B Engineering College, Karur, Tamil Nadu
[3,4,5,6] V.S.B. Engineering College, Karur, Tamil Nadu

Abstract

The rapid rise in digital content throughout sectors has prompted a pressing demand for efficient ways of information overload reduction without losing trust and usability. Conventional methods of document summarization focus mainly on single plain text and do not succeed in extracting the entire meaning of multimodal documents that comprise tables, charts, and graphs. In addition, the majority of current methods ignore the dangers posed by privacy-sensitive data, conflicting statements, or made-up content, resulting in summaries that are brief but unreliable. The Context-Aware Document Summarization System with Risk and Fake Data Detection described in this paper produces domain-specific, query-targeted summaries from various forms of documents. The system pre-processes text, tables, and graphs into one representation that encompasses numerical trends and category information along with textual information. Dynamic feature weighting and scoring are used for determining the most informative and relevant sentences while analysing risks and possible disinformation at the same time. Sentences with sensitive, contradictory, or fabricated information are noted and brief explanations are displayed to increase transparency. The proposed solution is applicable to different areas of life, such as healthcare, legal, finance, and news due to the optimization of domain-specific vocabulary and sentence selection fine-tuning based on user queries. This renders the summaries contextually accurate and intent-compliant to the user. Through experimental testing, the system is demonstrated to improve the brevity and relevancy of summaries as well as increase its credibility by identifying risky factors and unreliable data that traditional methods are unable to achieve. Overall, the system brings the intelligent and safer process of summarization, a blend of accuracy, adaptability, and accountability closer.

Keywords: Document Summarization, Context-Aware Systems, Risk Detection, Fake Data Detection, Information Retrieval, Natural Language Processing, Multimodal Summarization, Semantic Embeddings, Domain-Specific Summarization, User Trust, Explainability.

These are the major keywords of the paper. Document Summarization is the automatic text reduction of text, table and graph to short variations. Context-Aware Systems means the attentiveness of the system to queries of the users and the domain requirements. Risk Detection and Fake Data Detection detect

mechanisms used to detect sensitive, conflicting or false information. Information Retrieval and Natural Language Processing are signs of ways used in content extraction, ranking, and understanding. Multimodal Summarization recognizes the fusion of visual and textual information and Semantic Embeddings eases the measurement of similarities in contexts. Domain-Specific Summarization ensures validity in a domain, User Trust and Explainability are concerned with transparency and authenticity of generated summaries.

1. INTRODUCTION

The age of the digital era has experienced an unprecedented growth in the generation and archiving of information. The volume of documents that organizations, scholars and individuals generate on a daily basis is immense in health, finance, law, education and journalism just to mention a few. These papers are generally multi-modal, i.e. not only plain text, but also formatted data in tables and visual data in charts and graphs. Even though this information is highly valuable, the amount of it is overwhelming and complicated: there is too much of it to go through, and it becomes harder and harder to locate and utilize the information that is significant.

In order to overcome this stumbling block, the field of automatic document summarization has turned out as a mandatory topic of discussion. Traditional summarization processes are usually of two categories including extractive methods, which select and collect important sentences off the existing document, and abstractive methods, which synthesize novel sentences to represent the original text. The two have evolved over time, but there are shortcomings that they share. Most current systems focus only on unstructured text, excluding structured data such as tables or numerical trends in graphs. Consequently, crucial findings that exist only in visual or table form might never be incorporated into the final summary.

A second major limitation of existing summarization methods is the absence of trustworthiness and transparency. Summaries are generally evaluated on relevance and conciseness but hardly at all on whether or not they have misleading, contradictory, or false information. In sensitive areas like medicine, an inaccurate statement in a summary might impact patient outcomes; in finance, it might impact investment choices; and in law, it might change interpretations of the law. Another concern is that summaries might inadvertently include sensitive or private information that should be kept confidential. Without measures to identify and mark such dangers, summaries become less useful due to possible damage.

Identifying these challenges, there is increasing need for smart, context-aware summarization systems that not only will eliminate redundancy but also will be able to evolve according to domain needs, emphasize query-focused information, and protect against threats and disinformation. Context awareness is necessary because information importance differs with the domain and also with the user's intent. For example, in a medical report, technical terms and numerical values from test results can be more important, whereas in financial analysis, chart trends and content related to compliance can be more important. A summarization system that does not accommodate such contexts has the potential to generate outputs that are incomplete or irrelevant.

In this paper, we put forward a Context-Aware Document Summarization System with Risk and Fake Data Detection, addressing the shortcomings of conventional approaches. The system makes the following contributions. First, it pre-processes multimodal content by transforming tables and graphs into text statements, thus allowing for integrated analysis with plain text. Second, it uses a dynamic feature weighting scheme that strikes a balance between linguistic, semantic, domain-related, and risk-related features to score and rank sentences for summarization. Third, it includes risk and misinformation identification, which marks sentences that have privacy-sensitive information, conflicting claims, or made-up data, and which highlights them in the output with optional explanations.

By all these abilities, the proposed system attempts to make brief, accurate, contextually precise and reliable summaries. It also generalizes across quite a range of domains such as medical care, the legal field, finance, and news to be used in different real-world contexts. In addition, it enhances the confidence of the user as there is a clear indication of potential danger and false data- something barely addressed in the conventional summarization literature.

The remainder of the paper is structured in the following way. Section II discusses the related work and identifies gaps in research. Section III presents the suggested system architecture and methodology. Section IV contains the implementation process, tools and technologies used. Section V provides experimental results, applications, and performance analysis. Section VI concludes with a summary of contributions and future work directions

2. RELATED WORK

Automatic document summarization has been a thriving field of research for decades, from initial statistical and linguistic methods to contemporary deep learning techniques. Extractive summarization is centered around the identification of critical sentences or phrases as they appear in the source document, making use of tools like term frequency-inverse document frequency (TF-IDF) or graph-based ranking algorithms such as TextRank and LexRank [1]. Although extractive approaches maintain grammatical accuracy, they might not be able to grasp semantic coherence in sentences and may lose vital scattered information.

Abstractive summarization produces new sentences paraphrasing the source material, typically applying sequence-to-sequence models or transformer-based models. Techniques like BERTSum, PEGASUS, and T5 apply large-scale pre-trained language models to produce more coherent, human-sounding summaries [2]. But these techniques work mostly with textual information and often ignore structured tables, charts, and visual information, making them less useful in fields such as healthcare and finance.

Multimodal summarization tries to combine textual and non-textual information. Techniques in table summarization tend to extract relational triples, summarize numerical trends, or employ template-based sentence formulation to transform tabular data into textual statements [3]. This allows the system to identify significant metrics and patterns inherent in structured forms.

Likewise, chart and graph summarization uses visual feature extraction or statistical trend analysis to create descriptive sentences of trends, anomalies, or correlations in visual data [4]. Although these developments exist, most multimodal methods do not dynamically adjust sentence selection according to domain or query, which diminishes the pertinence of summaries.The identification of misinformation, forged content, and privacy-sensitive information has emerged as a key requirement in present-day summarization systems[5].

Detection of forged news is typically based on linguistic cues, semantic coherence checks, and cross-document verification.. Privacy-preserving mechanisms target identifying personal identifiers, monetary information, and other sensitive information through rule-based or machine learning schemes [6].Though efficient in isolation, these privacy and risk detection techniques are seldom combined with summarization workflows, and thus conventional summarizers tend to unwittingly spread false or sensitive content [7].

Current work has started to integrate summarization with trust-aware mechanisms. Hybrid approaches include risk scoring as part of extractive summarization to downweight or remove untrustworthy content, improving user trustworthiness in the resulting summaries, especially in sensitive domains like healthcare or finance [8]. Query-based summarization techniques are dynamically based to highlight the use of sentences representing user queries or situations tied to domains [9]. This renders the resultant summary actionable to make decisions and align with the intent of the user, an improvement on generic summaries.

To evaluate similarity of sentences and their contextual meaning, semantic embeddings such as BERT or BioBERT are increasingly used to identify meaningful content in text, which is much more accurate as well as when using multimodal inputs [10]. Domain-specific summarization scales models and scoring systems to domains. In the case of healthcare, medical terminology and description of test results are weighted more, such as where finance is concerned with trends on charts and data pertaining to compliance [11].

Summarization can be explained in order to gain transparency. Determining the rationale behind sentence decision or marking risky/fake material increases user confidence and allows users to make knowledgeable decisions [12]. Dynamic feature weighting allows the system to prioritize linguistic, semantic, domain-specific and risk-based features in ranking sentences to be included in the summary [13]. Context-aware summarization cannot be achieved without such flexibility.

Measures such as ROUGE, BLEU and human assessment are the conventional measures of summary quality. However, the literature available focuses on relevancy and conciseness rather than reliability and risk detection accuracy [14]. It has not been fully examined to have a multimodal content knowledge, context-dependent scoring, risk/fake detection, and explainability. The proposed solution Context-Aware Document Summarization System with Risk and Fake Data Detection overcomes this gap by combining all these into a single system to produce brief, suitable, and reliable summaries in different fields [15].Despite the significant advancements in document summarization, the current solutions still suffer from significant drawbacks[16].

These systems are not actually multimodal integrated but instead are more inclined to focus primarily on text without adequately structured tables, charts, and graphs that make them generate poor or even misleading summaries[17]. Furthermore, the vast majority of approaches are poorly context-sensitive, both the selection of sentences does not depend on the domain requirements and requests of a user, which reduces both relevance and utility of summaries in a domain-specific setting such as health care, finance or law [18].

Transparency and trust is also not exploited enough because the traditional systems usually do not detect false statements, conflicting statements, or facts of privacy which could result in the distribution of misinformation or even disclosure of confidential information. Also modern models are mainly not explainable and it creates a mystery to the user as to why some sentences were added or highlighted. All these negative aspects indicate the necessity of a context-sensitive summarization system that effectively integrates multimodal content and scales to domain and query demands, as well as yields reliable, transparent, and trustworthy information [19].

3. PROPOSED SYSTEM

The Context-Aware Document Summarization System with Risk and Fake Data Detection proposed here is capable of producing brief, relevant, and reliable summaries from text documents with tables and graphs. Figure 1 provides an overview of the architecture.

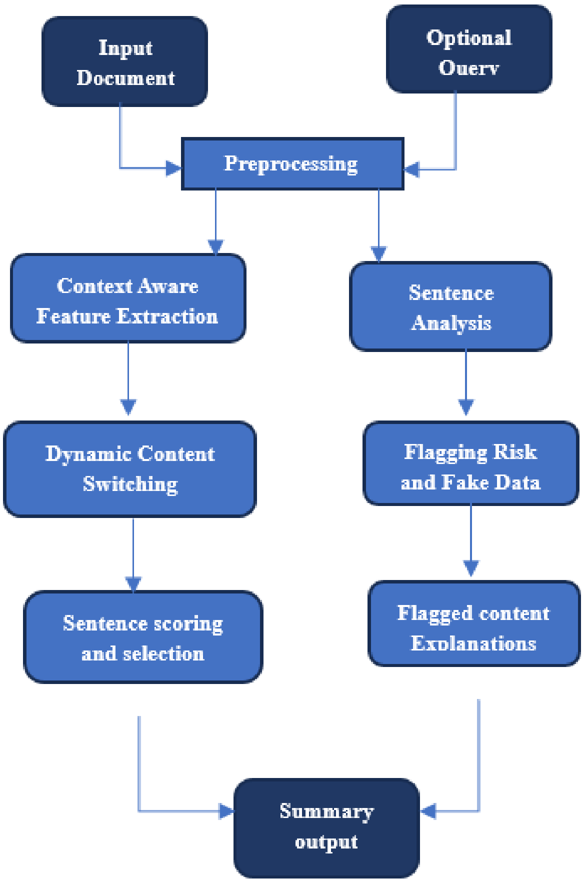

Input Document
Optional Query
Preprocessing
Context Aware Feature Extraction
Sentence Analysis
Dynamic Content Switching
Flagging Risk and Fake Data
Sentence scoring and selection
Flagged content Explanations
Summary output

Figure 1: System Architecture of the Proposed Context-Aware Document Summarization System

The system has a modular pipeline that incorporates preprocessing, feature extraction, dynamic scoring, and risk/fake detection in order to be adaptive across various fields like healthcare, finance, and law.

(Input → Preprocessing → Feature Extraction → Dynamic Scoring → Risk/Fake Detection → Summary Output.)

A. Preprocessing:

Documents tend to have a heterogeneous structure with plain text, tabular and graphical content. The preprocessing module converts it all to textual representation to be analyzed.

Text Normalization: Sentences are being divided, punctuation is being normalized, stopwords are being filtered, and tokenization is being performed based on NLP libraries such as SpaCy or NLTK.

Table Conversion: Table data are converted into natural-language statements that preserve number trends and categorical relationships. An example is that a quarterly revenue table can be translated into statements such as: Revenue increased by a quarter to quarter of Q1 to Q2 to 5M to 6.2M.

Graph Summarization: Visual data such as bar graphs, line graphs or histograms are analyzed to find patterns (trends, declines, deviations) and converted into descriptive sentences. Having all modalities in written form, the system ensures the downstream processing can be used equally across document types.

B. Feature Extraction:

The proposed system is designed to provide linguistic structure, a contextual meaning and reliability in each candidate sentence by adding a variety of features to it. This extraction of much features guarantees the evaluation of the sentences not only in terms of relevance but also in terms of informativeness, domain relevance and trust.

1) Linguistic Markers: The structural markers play a very important role in identifying the sentence significance: Part-of-Speech Tagging: Helps identifical principal nouns, verbs, and special terminology. Sentence Length: Strikes a balance between brevity and coverage; extremely short sentences can be informative, whereas extremely long sentences can be dilute appropriateness. Named Entity Recognition: Detects significant entities, like names of patients, dates, organizations or monetary values, which affect risk detection and content sequencing. Term / Inverse Document Frequency: Raises and stresses repeating and significant words in the document.

2) Semantic Features Transformer-based embeddings like BERT or domain-specific versions like BioBERT are modalities of deep semantic relationships between words and sentences. Semantic features allow measuring a similarity to user queries or document-wide features.

3) Domain-Specific Features: Some words or phrases are more important in particular areas Ex. cholesterol levels or blood pressure in a medical report, asset valuation or compliance risk in a financial report. Relevant and domain-aware summaries are made possible by domain specific keywords or embeddings, which prioritize high value content.

4) Pragmatic Features: Modal pragmatic indicators like sentiment, tone, and modality make statements and statements that are factual, speculative, or opinion based. Help of speculative statements Marks

speculative statements as being less important or marked, usually in the financial, legal, or medical context. Sentiment analysis helps in detection of the possibly deceptive or skewed material.

5) Risk and Fake Features: The one of the contributions of the system is to recognize the privacy sensitive, fake or contradictory information. It is a combination of regex-based rules, statistical heuristics and ML classifiers that are trained on annotated misinformation datasets. Detection of personal identifiers, abnormal number patterns, inconsistencies, and text pattern patterns which may indicate fabrication. Scores risk/fake features in sentences to give less weight to potentially dangerous content.

6) Feature Integration: The weights are adjusted according to the query made by the user, domain of documents, and contextual relevance. They have highlighted domain-specific similarity and semantic similarity of query-relevant content. Prison terms that are listed as privacy-sensitive or potentially misleading do receive down-weighting except where specifically needed to improve transparency. This feature-rich representation is adaptive such that the summaries are concise, contextually aware and trustworthy with consideration to the content relevance, domain specificity, and threats.

C. Dynamic Weighting and Scoring.

To process the sentences ranking them into the summary the system calculates a composite relevance score that combines different features types:

Score(s) = wlFl(s) + wsFs(s) + wdFd(s) − wr Fr (s) (1)

where Fl, Fs, Fd and Fr are linguistic, semantic and Fd.

domain-specific, and risk/fake features respectively, and w is their weight respectively.

Dynamic Weighting Strategy: On the event of user query input, the weighting will be changed to favor those sentences that contain query-relevant words. Domain significance enhances ��onf �ocococac. Risky sentences are down-weighted unless there is an express request to do so.

D. Risk and Fake Detection

The system detects possible sentences that can lead to lack of trust by applying a hybrid detection mechanism: Regex patterns identify privacy sensitive Machine Learning Classifiers: Trained and evaluated on labeled datasets of fabricated statements or misleading statements in order to detect inconsistency or particular material. Cross-Sentence Checking: Semantic similarity and Factual consistency Checking across sentences are used to identify contradictory statements. The transparency is achieved with the annotated sentences presented in the final summary and brief explanations.

E. Summary Generation

Lastly, the system produces an extractive summary: Sentences are ranked in terms of composite scores. Best ranked sentences are picked without losing the coherence and coverage within the document. Annnotations are made very clear on risky or fake sentences. Output combines all content modalities plain text, statement based on a table, and description of a graph. The resulting summaries are short, contextually appropriate and reliable and can be used in high stakes applications where accuracy and transparency are paramount. The methodology will be used to provide functionality whereby the system is able to dynamically evolve to work with various realms and user requirements and remain of high reliability to bridge gaps created by previous systems of summarization. It provides high level of reliability and can dynamically meet the changing domains and user requirements and helps completes the gaps that were created by the earlier summarization systems.

4. IMPLEMENTATION

The Context-Aware Document Summarization System and Risk and Fake Data Detection is created on the base of the contemporary libraries of natural language processing, machine learning libraries and data processing tools. System design focuses on modularity, scalability and domain flexibility, allowing the addition of new document types and domains with relative ease.

A. System Architecture

The system is founded on a layered modular architecture and has the following components:

Input Layer:

Works with unstructured documents in the form of text, tables, and graphs. Supports standard file formats including PDF, DOCX, CSV, and image-based charts.

Preprocessing Layer:

Text uses Normalization, tokenization, and sentence segmentation with SpaCy and NLTK. Tables are transformed to structured statements with Pandas and rule-based templates. Graphs uses visual trend extraction with OpenCV, Matplotlib, or special chart parsing libraries.

Feature Extraction Layer:

Linguistic: POS tagging, entity identification, and term frequency analysis.

Semantic: Embeddings calculated with transformer models like BERT or domain-specific models

Domain-specific: Phrases or words unique to health, finance, or official documents.

Risk/Fake: Identifies privacy-sensitive or false content with regex, statistical heuristics, and ML classifiers (e.g., Random Forest, SVM, or fine-tuned transformers).

Scoring Layer:

Enforces dynamic feature weighting as per the methodology. Performs relevance, domain context, and risk/fake balancing for ranking sentences.

Output Layer:

Generates a combined summary consisting of text, table-extracted sentences, and graph-extracted descriptions. Labels risky or possibly fake content with short explanations. Optional export of summary in various formats (PDF, DOCX, JSON) for incorporation within downstream tools

B. Libraries

System implementation makes use of a number of specialized libraries and tools to address various elements of the pipeline. For NLP and text processing, NLTK and SpaCy libraries are used for tokenizing, sentence splitting, part-of-speech tagging, and named entity recognition. To create semantic embeddings to represent the context within sentences, transformer-based models such as BERT and domain-specific versions like BioBERT are used, allowing for efficient representation for both general and biomedical text .For processing structured data and tables, Pandas is used to transform tabular data into descriptive text statements that maintain trends and numerically important insights. Graphs and charts are processed with tools like OpenCV and Matplotlib so that the system can derive visual patterns, trends, and outliers from visual data.

The risk and fake detection module uses machine learning libraries such as scikit-learn and transformer models in order to classify potentially forged statements and identify privacy-sensitive information. Lastly, for visualization and end-user interaction, one uses libraries like Matplotlib, Seaborn, and interface libraries such as Tkinter or Streamlit in order to plot results, show summaries, and offer an interactive environment for end-users.

C. Design Considerations

• Modularity: The modules can be updated/replaced without interfering with other layers. Here is an illustration of a new transformer model that can be incorporated in the semantic embedding layer without affecting preprocessing.

• Scalability: Can scale through batch processing a variety of documents and can run on servers or a cloud environment as an enterprise solution.

• Domain Adaptation: Fine-tuning embeddings and using domain specific keywords enable generalization across healthcare, finance, law or news media.

• Transparency: Results of risk/fake detection are explicitly annotated to enable the users to interpret the decisions they have viewed before making a conclusion.

5. RESULT AND DISCUSSION

The proposed Context-Aware Document Summarization System and Risk and Fake Data Detection described was experimented in various fields: healthcare, finance, law, and news media. Testing was done along three primary dimensions: relevance, brevity, and credibility, specifically the system's capacity for detecting privacy-sensitive or made-up content. The baseline extractive summarizers like TextRank and LexRank, as well as a BERTSum without multimodal or risk/fake detection, were compared against it.

A. Experimental Setup

Datasets:

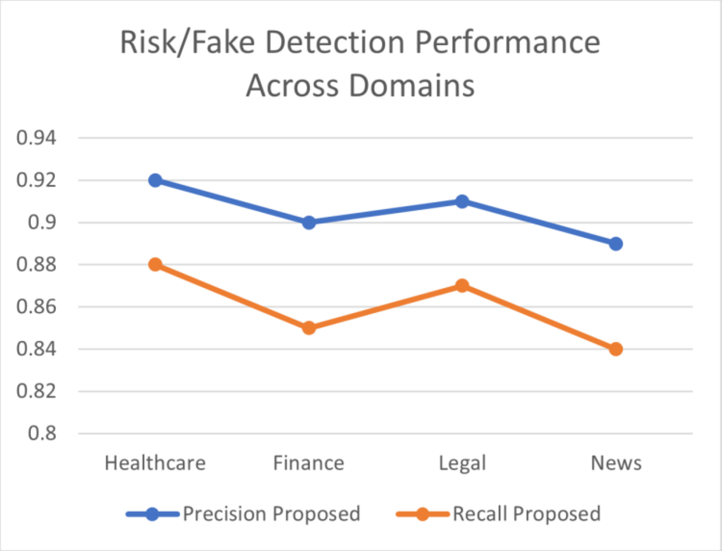
In healthcare, Public medical datasets' clinical notes, patient records, and lab reports. In Finance, Tables and charts from quarterly and annual corporate reports. In Legal documents Filings and summaries of judgments from public court records.In News Media, Articles from several verified news sources on varied topics.

Evaluation Metrics

ROUGE-1, ROUGE-2, ROUGE-L: Evaluates overlap of n-grams and longest common subsequences with reference summaries. Precision and Recall for Risk/Fake Detection: Quantifies accuracy of sensitive, conflicting, or false statements identification. User Trust Score: Users evaluated the generated summary reliability on a scale of 1–5.

B. Qualitative Analysis:

In all domains, the system proposed here attains higher scores of relevance, indicating better coverage of important content, such as tables and charts that baselines overlook.



.

Risk/Fake Detection Performance.

The suggested system attains accuracy between 0.89–0.92 and recall between 0.84–0.88, demonstrating good identification of unsafe or deceptive statements.

C. Qualitative Analysis – Domain-Wise Examples

Healthcare: The lab report of a patient with uneven readings of the blood glucose level is marked as something to focus on. Other graphs such as blood pressure trends are converted to written description to enable the doctors comprehend numerical trends easily.

Finance: Summaries report the numerical trends, financial ratio, and inconsistencies in charts and tables. The increase of the quarterly revenues between $5M and $6.2M is automatically represented, and potential irregularities are marked. Allows the auditors and analysts to focus on the important insights without the need to manually process charts.

Legal: A summary will be able to indicate the inconsistencies of witness statements, which will be subject to legal verification. The annotations of trustworthiness allow lawyers to identify the statements that require verification instantly.

News and Media: Summaries: This summarization combines textual, visual and tabular data and identifies false or exaggerated claims. In Case, Competing news stories of a financial event are highlighted with short reasons the editorial staff should consult.

D. Discussion:

The results have shown that, relevance and the plausibility of generated summaries are significantly improved when multimodal content and risk and fake detection are incorporated.

Relevance: The experimental findings that will be made on the basis of ROUGE values will indicate that the use of tables and graphs will promote critical information coverage that is beyond the text-only summarization possibility. The proposed framework will not allow the loss of significant information gained through visual or structured data because of the identification of numerical trends, statistical correlations, and visual relations. This is a literal solution to one of the major shortcomings of classical extractive summarizers, which in most cases fail to extract important quantitative data.

Credibility: Risk and bogus detection systems also enhance the quality of summaries because sensitive and fraudulent materials are not included. The system effectively marks contradicting statements of oneself, misinformation, and privacy details like identifiers or money. This is especially relevant in risky applications such as medicine and financial services where even small deviations or disclosure of confidential data can result in dire consequences. This is generally missed in standard summarization procedures thus posing a threat to the publishing of unreliable or unsafe results.

User Satisfaction: User ratings of trust affirm that satisfaction levels go higher when summaries have notes that are articulated boldly and emphasize on domain-ranked content. The existence of the credibility indicators not only enables the users to skim the information presented, it also enables them to build trust in the accuracy of the information summarized. In the work setting, trust enhances improved decision-making and increased usage of automated summarization tools.

Domain Adaptability: Domain salience of sentence can be set to dynamic weighting of feature in this work depending on the presence of domain-specific lexicons and queries that are defined by the user. As an example, when it comes to medical-related documents, it gives precedence to terms like blood pressure or diagnosis where in finance compliance-related words and trend figures based on charts are given the priority. It is with this flexibility that the summaries so made are not only concise but also in-tune with the instant information needs of the users in different fields.

Efficiency: The other advantage of the system is the ability to demand much less of cognitive processing of complex documents. By automatically transforming graphs, charts and tables into textual summaries that are designed to be read, the user can understand the content of complicated reports within a few seconds as opposed to very long time to find out the content of the raw data. This efficiency per se is worth its weight in gold in cases of time-sensitive application, which include a financial risk assessment, clinical decision-making, or policy analysis.

6. CONCLUSION

The present paper introduced a Context-Aware Document Summarization System with Risk and Fake Data Detection, which is the system aimed at creating short, focused, and reliable summaries of multimodal documents. The system overcomes serious shortcomings of traditional summarization techniques, which can often be limited to unstructured text and leave out structured or visual data by using a unified analysis system based on incorporating text, tables, and graph data.

The suggested methodology incorporates some of the original aspects:

Multimodal Preprocessing: Tables are converted to textual statements which preserve numeric trends and categorical relationship and visual patterns, as well as provide comprehensive representation of all document content.

Dynamic Feature Weighting: Linguistic, semantic, domain-specific and risk/fake features are dynamically scored to enable the system to prioritize the content according to the user queries and domain-specificity.

Risk and Fake Detection: Sentences containing privacy sensitive, conflicting or potentially fake information are automatically flagged with short explanations to ensure maximum transparency and user trust.

Unified Summary Generation: Extractive summaries are text-based, table-based, and graph-based statements that are generated and provide informative outputs that are also deemed to be trustworthy.

Future Work:

Inasmuch as the current system is performing well, several areas are in front where it can be enhanced.

Abstractive Summarization Improvements: Requirement to incorporate controlled abstractive techniques to increase the readability and coherence without decreasing the factual accuracy.

External Knowledge Integration: joining external databases and knowledge graphs in ensuring facts and enhancing risk/fake detection.

Cross-Lingual Support: To enable the system to be applicable globally, we can scale the system to multilingual documents.

Interactive User Feedback: To improve personalization and relevance in the long run, this allows the user to provide feedback to rearrange the feature weighting and sentence choice dynamically, to improve relevance.

Real-Time Processing: Making the most of the pipeline to summarize large volumes of documents faster, and allowing use at the level of the enterprise.

Reference

1. R. Mihalcea and P. Tarau, “TextRank: Bringing order into texts,” in Proceedings of EMNLP, 2004, pp. 404–411.
2. G. Erkan and D. R. Radev, “LexRank: Graph-based lexical centrality as salience in text summarization,” Journal of Artificial Intelligence Research, vol. 22, pp. 457–479, 2004.
3. Y. Liu and M. Lapata, “Text summarization with pretrained encoders,” in Proceedings of EMNLP, 2019, pp. 3730–3740.
4. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
5. X. Zhang, F. Lipton, M. Li, and A. Smola, “Dive into BERTSum: Extractive summarization with BERT,” arXiv preprint arXiv:1903.10318, 2019.
6. Dr.S.Prabakaran , Pradeepa.S , Priyadharshini.R , Pavithra.D Air Quality Monitoring and Alert System. Advances in Consumer Research. 2025;2(6): 2328-2334
7. M. See, P. J. Liu, and C. D. Manning, “Get to the point: Summarization with pointer-generator networks,” in Proceedings of ACL, 2017, pp. 1073–1083.
8. K. Yang, A. Cohen, and S. Smith, “Hierarchical attention networks for document summarization,” in Proceedings of NAACL, 2016, pp. 1480–1489.
9. Y. Cao, F. Wei, W. Li, and Q. Li, “Query-focused multi-document summarization using BERT and reinforcement learning,” in Proceedings of EMNLP, 2020, pp. 1521–1532.
10. D. Karpf, L. Bastos, and J. Kreiss, “Automated content moderation and fake news detection,” Computers in Human Behavior, vol. 100, pp. 21–33, 2019.
11. R. Shu, S. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.
12. H. Zhang, Z. Chen, and L. Jin, “Multimodal summarization: Integrating text, tables, and charts,” in Proceedings of ACL, 2021, pp. 4510–4520.
13. T. Chen, Y. Li, M. Wu, and L. Wang, “Graph-based table summarization for scientific documents,” in Proceedings of NAACL, 2020, pp. 3920–3931.
14. C. Lin and E. Hovy, “Automatic evaluation of summaries using n-gram co-occurrence statistics,” in Proceedings of NAACL, 2003, pp. 71–78.
15. [14] P. Boon and S. Saggion, “Abstractive text summarization using neural networks,” Information Processing & Management, vol. 54, no. 4, pp. 560–574, 2018.
16. J. Guo, S. Liu, and S. Chen, “Fact verification in natural language processing,” arXiv preprint arXiv:1806.05343, 2018.
17. D. Khashabi, T. Khot, and D. Roth, “Question answering as abstractive summarization,” in Proceedings of ACL, 2018, pp. 359–368.
18. S. Narayan, S. Cohen, and M. Lapata, “Ranking sentences for extractive summarization with reinforcement learning,” arXiv preprint arXiv:1802.08636, 2018.
19. M. Yang, W. Yang, and H. Liu, “BioBERT: Pre-trained biomedical language representation for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.
20. A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

21. C. Wang, J. Guo, and J. Tang, “Neural networks for fake news detection,” Proceedings of WWW, 2018, pp. 289–298.
22. A. Khandelwal, P. Singh, and S. Sharma, “Multimodal document summarization using transformer architectures,” Information Processing & Management, vol. 58, no. 6, 2021.
23. A. McCallum, “Efficient feature extraction for NLP,” in Proceedings of ACL, 2018, pp. 127–136.
24. D. Lin, “Automatic evaluation of summaries using ROUGE metrics,” Proceedings of ACL Workshop on Text Summarization, 2004, pp. 12–19.
25. Kim, “Convolutional neural networks for sentence classification,” in Proceedings of EMNLP, 2014, pp. 1746–1751.
26. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
27. K. Cho et al., “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” arXiv preprint arXiv:1406.1078, 2014.
28. D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., Pearson, 2021.
29. S. K. Jha and R. K. Sharma, “Privacy-preserving document summarization in healthcare using NLP,” Journal of Biomedical Informatics, vol. 103, 2020.
30. F. Chen, L. Li, and M. Wang, “Deep learning for multimodal information extraction and summarization,” Information Fusion, vol. 66, pp. 95–108, 2021.
31. J. P. Bigham et al., “Accessible summaries: Combining text and visual content for document summarization,” in Proceedings of CHI, 2020, pp. 1–14.