Health Misinformation Detection: A Chunking Strategy Integrated to Retrieval-Augmented Generation

Walid Taib[1,*,†], Idriss Saadallah[2,†], Abdelali Ichou[2], Abderrahmene Hamdi[3], Alessandro Bruno[4], Pier Luigi Mazzeo[5], Aladine Chetouani[6], Marouane Tliba[2] and Mohamed Amine Kerkouri[2]

[1]*Higher National School of Telecommunications and Information Technologies and Communications, Es Senia Oran, Algeria*
[2]*University of Orleans, Orleans, France*
[3]*Higher National School of Computer Science, Oued Smar Algiers, Algeria*
[4]*IULM University, Department of Business, Law, Economics, Consumer Behaviour - "Carlo A. Ricciardi", Via Carlo Bo 1, Milan, 20143, Italy*
[5]*ISASI Institute of Applied Sciences and Intelligent Systems-CNR, 73100 Lecce, Italy*
[6]*University Sorbonne - Paris Nord, Villetaneuse, France*

Abstract

Generative AI (GenAI) and Natural Language Processing (NLP) have advanced significantly in recent years, exhibiting breakthroughs and pushing the bar of accuracy rates in text mining. Cascade effects have been observed in many application domains, spanning text analysis, question answering, classification, and new textual content generation. The latter has allowed many end-users to perceive AI as ready-to-go solutions to optimise their daily workflow. However, dark and bright sides lurk behind textual content generation, as trustworthy and unverified content can be effortlessly generated. That has fuelled a significant challenge in our society: fake news. Although fake news has existed for a while, it remains an unsolved issue. Generative AI has brought it to a new level by enabling the automated production of large volumes of high-quality, individually targeted fake content. Our work is part of the HeReFaNMi (Health-Related Fake News Mitigation) project, which focuses on health-related fake news mitigation by using NLP, Language Models, and a Retrieval-Augmented Generation (RAG) system. We propose a new chunking mechanism that streamlines the overall RAG framework pipeline. BERT and BERT+RAG have been compared on the health-related fake news classification task on a dataset of 2000 health-related articles equally split into two categories ('fake' and 'credible'). Preliminary experimental results reveal improvements in Accuracy, Recall, and F1-score.

Keywords

Generative AI, Natural Language Processing, Fake News, Health Misinformation, Language Models, Retrieval-Augmented Generation, Misinformation Detection

1. Introduction

The advent of language models has changed the field of NLP and artificial intelligence making a meaningful contribution to diverse domains, going from creating human-like text to powering chatbots. These models have proven an impressive capability to handle difficult tasks like translation, summarization, and even creative writing [1].

Language Models [2] fuel popular platforms like ChatGPT and Gemini. In particular, they rely upon the Transformer architecture introduced in the seminal work "Attention is All You Need." [3]. Transformers have impressively improved how textual information is handled with self-attention and

AIxPAC: Workshop on Artificial Intelligence for Perception and Artificial Consciousness, November 25–28, 2024, Bolzano, Italy
*Corresponding author.
†These authors contributed equally.
walid.taib@ensttic.dz (W. Taib); idriss.saadallah@univ-orleans.fr (I. Saadallah); abdelali.ichou@etu.univ-orleans.fr (A. Ichou); Ka_hamdi@esi.dz (. Hamdi); alessandro.bruno@iulm.it (A. Bruno); pierluigi.mazzeo@cnr.it (P. L. Mazzeo); aladine.chetouani@univ-orleans.fr (A. Chetouani); marouane.tliba@univ-orleans.fr (M. Tliba); mohamed.a.kerkouri@gmail.com (M. A. Kerkouri)
0000-0001-5191-1662 (A. Bruno); 0000-0002-7552-2394 (P. L. Mazzeo); 0000-0002-2066-4707 (A. Chetouani); 0000-0002-3178-3509 (M. Tliba); 0000-0002-7479-6879 (M. A. Kerkouri)
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-3923/Paper_5.pdf

attention principles. These represent a breakthrough in generative AI, allowing predicting the next word in a sequence [4]. However, despite their potential, Language Models face many challenges that cause them to miss a true understanding of the text they generate [5]; one of the most popular side-effects goes under the name of hallucinations [6] , where the model produces incorrect information that seems plausible [7] [8]. In some areas, reliability and accuracy are paramount, with healthcare probably on top of the stack. For instance, relying on a model that may generate false or unverified information may lead to serious consequences.

The ability of Language Models to generate human-like text makes them a powerful tool, but it also makes them susceptible to misuse [9]. Zooming in on healthcare, many fake news articles spread over the Internet during the COVID-19 pandemic [10] , causing a concerning lack of trust towards national healthcare systems worldwide [11]. For that reason, we are in need of reliable systems, which can effectively leverage the power of Language Models while mitigating risks [12]. Unlike traditional Language Models that rely solely on internal model parameters, RAG (Retrieval Augmented Generation) [13] systems allow the retrieval and incorporation of data from external sources. That represents a mitigation solution in healthcare as authentic and trustworthy external health-related sources provide reliability to the system[14]. Furthermore, the RAG-based approach improves the model's performance and significantly reduces the likelihood of generating hallucinations or inaccurate responses[15].

This paper presents a new contribution through the integration of a new chunking mechanism into the RAG framework [16]. The chunking system allows for the processing of longer and more complex text by breaking them into smaller and interpretable chunks that fit the context window size [17, 18], and sending them to the Language Models as context. Our approach also leverages the power of prompt engineering to fine-tune the model's inputs for more accurate.

Our contribution represents a meaningful effort to adapt and improve RAG systems to combat healthcare-related fake news. Through a retrieval-based system, we aim to provide a robust solution to the growing problem. This paper will explore the technical aspects of our system and the potential of our approach to mitigate these problems effectively through an improved novel chunking strategy.

2. Related Work

2.1. Chunking in Language Models (LMs)

When building applications that rely on natural language processing (NLP) tasks, such as semantic search or document summarization, one of the most critical aspects is ensuring that the text is represented in a way that maintains its meaning and relevance. Chunking is an essential technique in this regard, particularly when working with Language Models that have token limits. Previous work has explored various chunking techniques to handle these limitations effectively [19, 20].

Fixed-size Chunking is the most straightforward approach, where the document is split into equally sized segments based on a predefined number of tokens or characters. Fixed-size chunking is easy to implement and works well in many scenarios [20]. However, this approach risks losing semantic context if the division happens mid-sentence or mid-idea, potentially reducing the effectiveness of Language Models in downstream tasks like retrieval and summarization [17].

Content-aware strategies focus on partitioning a document based on its inherent structure, such as sentences, paragraphs, or sections, allowing the system to preserve meaningful boundaries and enhance both retrieval and processing performance.

Sentence Splitting: This method ensures that each chunk consists of complete sentences, thus maintaining readability and coherence. Sentence splitting is especially advantageous in tasks like summarization and question answering, where preserving sentence-level integrity is crucial for maintaining context and meaning [21].

Recursive Chunking: This technique involves dividing documents using predefined separators (e.g., paragraph breaks, sentence boundaries) while ensuring that semantically relevant content is preserved. Recursive chunking is flexible, with parameters such as *chunkSize* (defining the maximum allowable chunk size) and *chunkOverlap* (controlling the degree of content shared between adjacent chunks).

This approach has been particularly effective for processing long documents, ensuring that semantic continuity is maintained within and across chunks, thereby improving document understanding and task performance [22].

Documents written in markup languages (e.g., Markdown, HTML, LaTeX) implement tagging systems that categorize and structure text into both semantic and syntactic units. This intrinsic formatting offers a unique advantage for advanced chunking techniques [23], specifically designed to handle semi-structured data while accommodating its flexible schema. By harnessing the structured metadata and hierarchical markers, these techniques ensure that meaningful chunk boundaries are preserved, thereby improving the efficiency and accuracy of downstream processes such as information retrieval, content summarization, and data extraction. This structured approach enhances the interpretability of the data and ensures a more coherent representation of complex document structures [24].

3. Proposed Method

In this work, we present a new approach to enhance the performance of the Retrieval-Augmented Generation (RAG) by changing how the articles are split and retrieved. Our proposed method encompasses the following key steps (see Figure 1):

- **Sentence Splitting:** the article is divided into paragraphs. This step allows a finer-grained retrieval process by which the model can work on smaller, more focused parts of information.
- **Cosine Similarity Calculation:** We calculate the cosine similarity between the sentence embeddings to find semantically similar sentences. Those with high cosine similarity scores are grouped together to form coherent content chunks. The cosine similarity between two sentence embeddings, s_i and s_j, is calculated as:

$$\text{cosine\_similarity}(s_i, s_j) = \frac{s_i \cdot s_j}{\|s_i\|\|s_j\|} \tag{1}$$

- **Chunk Creation:** Sentences with high similarity scores are merged into chunks, in other words, similar ideas are grouped into one chunk. These chunks become the primary retrieval units, containing semantically related content, which improves the precision of the retrieved context.
- **Re-Embedding of Chunks:** After the chunks are created, each chunk is re-embedded using a semantic model. This process captures a more meaningful representation of the chunk in the latent space, This chunk-based approach improves the retrieval process by making sure that the retrieved content is more focused and semantically coherent.

The chunk-based approach enhances retrieval by ensuring the retrieved content is more focused and semantically coherent. A more contextually relevant document can then benefit the generated output. The following subsections describe how these modifications impact the retrieval and generation components of the RAG model.

3.1. Retrieval

The retriever, $p_\eta(z|x)$, now operates over these newly created chunks. We maintain the Dense Passage Retriever (DPR) bi-encoder architecture. Still, instead of retrieving entire documents, the model retrieves top-K chunks based on the cosine similarity between the input query x and each chunk embedding z. The retrieval probability is formalized as:

$$p_\eta(z|x) \propto \exp(\text{cosine\_similarity}(x, z)) \tag{2}$$

where $\text{cosine\_similarity}(x, z)$ measures the semantic similarity between the query x and the chunk z. This ensures that only the most relevant chunks are selected for generation.

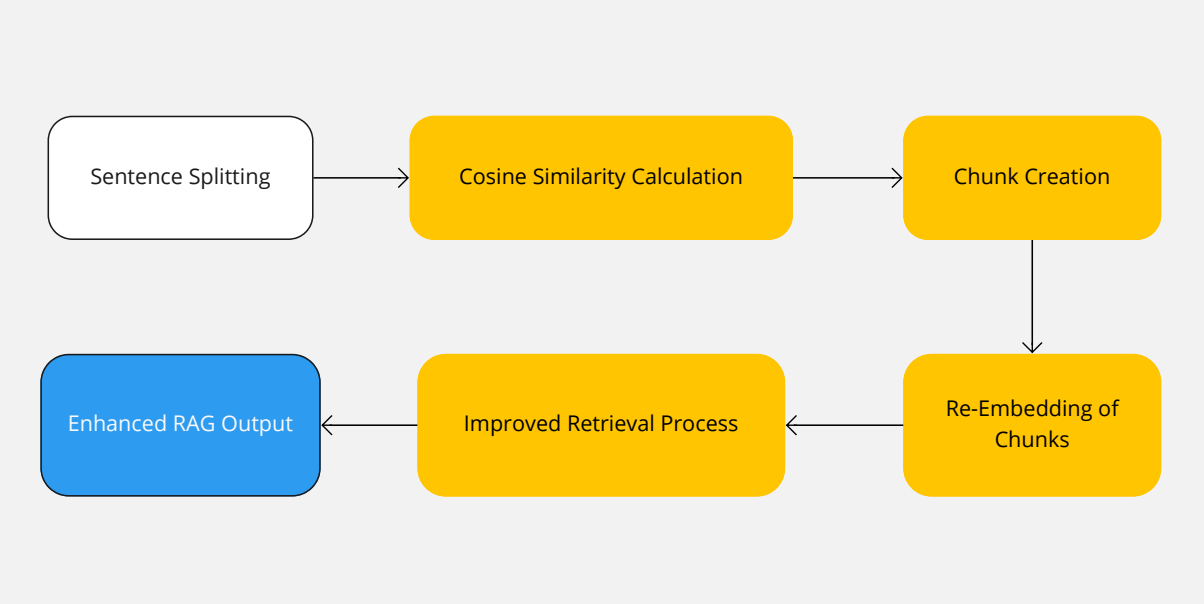

Figure 1: The diagram above graphically depicts the most meaningful aspects entailing the proposed method.

3.2. Generation

We employ two variants of the generation model: RAG-Sequence and RAG-Token [14], using the retrieved chunks as context to generate the target sequence y.

A single retrieved chunk z generates the entire output sequence in the RAG-Sequence Model. The probability of generating the sequence y is marginalized over the top-K retrieved chunks:

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-K}(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x,z) \tag{3}$$

The generator produces each token y_i based on the retrieved chunk and the previous tokens:

$$p_\theta(y|x,z) = \prod_{i=1}^{N} p_\theta(y_i|x,z,y_{1:i-1}) \tag{4}$$

Then, we implement RAG-Token to allow the model to select a different chunk z for each token. The probability of generating the output sequence y is defined as:

$$p_{\text{RAG-Token}}(y|x) \approx \prod_{i=1}^{N} \sum_{z \in \text{top-K}(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x,z,y_{1:i-1}) \tag{5}$$

The above-described approach is dynamic and allows the model to select the most relevant chunk for generating each token in the sequence.

By using cosine similarity to group similar sentences into chunks and re-embedding these chunks, we significantly improve retrieval precision. Each chunk contains semantically consistent information, making the retrieval process more focused and reducing irrelevant or noisy content. These modifications benefit both RAG-Sequence and RAG-Token models, leading to enhanced performance in sequence generation and classification tasks.

4. Results and Discussion

In this section, we present the results of our experiments comparing the performance of BERT[25] with and without the use of Retrieval-Augmented Generation (RAG). We employed the RAG-Token model

in our experiments over a dataset of 2000 news articles (equally balanced onto 'fake' and 'credible' categories) collected as part of the *HeReFaNMi* project [1], funded by NGI Search. The dataset, which is publicly available, was gathered using a web scraping framework and prompt engineering techniques, focusing on the classification of credible and fake news.

We evaluated the BERT [25] base model on this dataset in two scenarios: (1) without RAG, and (2) with RAG, where the model retrieves relevant text chunks before making classification decisions. The results of these experiments are summarized in Table 1:

Metric	Without RAG	With RAG
Accuracy	0.6575	**0.7010**
Recall	0.3116	**0.898**
F1 Score	0.4751	**0.767**

Table 1
Performance of BERT Base Without and With RAG

4.1. Performance Analysis

As seen in Table 1, BERT [25] without RAG achieves an accuracy of 65.75 percent, with a recall of 31.16 percent and an F1 score of 0.4751 percent. In contrast, when using RAG, the accuracy improves to 70.10 percent, while the recall increases significantly to 89.8 percent. This substantial improvement in both accuracy and recall when using RAG indicates that the model becomes more sensitive and effective in identifying relevant instances.

The F1 score, which balances precision and recall, is 0.4751 without RAG and increases to 0.767 with RAG. This improvement in the F1 score reflects that RAG not only enhances the model's recall but also results in better overall performance by improving precision, thereby boosting both recall and accuracy metrics.

4.2. Resource Efficiency and Model Comparison

One of the major advantages of using BERT[25], especially without RAG, is its efficiency in terms of computational resources. BERT [25] is lightweight and requires fewer resources to train and fine-tune than larger models such as LLaMA-3[26]. LLaMA-3, while powerful, demands extensive computational resources and large datasets, which are not always available or feasible to use. In contrast, BERT [25] can perform well even with smaller datasets and fewer computational requirements, making it a suitable choice for resource-constrained environments.

However, incorporating RAG into BERT [25] adds complexity to the model. Although RAG boosts recall, it introduces additional retrieval steps, which increase computational overhead. Despite this, RAG's ability to provide relevant context to the model helps enhance its performance, especially when fine-tuning the model on tasks requiring retrieval of external information.

4.3. Enhancing BERT with RAG

Our experiments show that BERT [25], when combined with RAG, becomes much better at recalling relevant information but at the cost of precision. This suggests that RAG is beneficial for cases where recall is critical (e.g., ensuring that all potential relevant information is retrieved), but further fine-tuning or balancing mechanisms are required to maintain high accuracy.

Future work could focus on optimizing the integration of RAG with BERT [25] to reduce the trade-off between recall and accuracy, perhaps by fine-tuning the retrieval process or introducing filtering mechanisms to prevent irrelevant information from being included in the classification process. Overall, while BERT [25] with RAG shows promise, particularly for recall-focused tasks, improvements can be made to enhance its overall performance.

[1]Health-care Related Fake NEWS Mitigation Project (HeReFaNMi)

5. Conclusion

In conclusion, the integration of Retrieval-Augmented Generation (RAG) into models like BERT [25] presents a promising approach for improving the detection of health-related fake news. By leveraging the enhanced retrieval mechanism provided by RAG, the system can access more accurate and contextually relevant external information, leading to better recall in identifying misinformation. Our results demonstrate the trade-off between recall and accuracy, with RAG significantly boosting recall at the cost of introducing more false positives. This underscores the importance of refining retrieval processes and balancing model precision and recall for practical applications. Future work could focus on optimizing this balance and further improving the resource efficiency of such models. Ultimately, our proposed method holds significant potential for combatting misinformation, particularly in critical areas such as public health, where accuracy is paramount.

6. Acknowledgments

The contribution is funded by the grant awarded for HeReFaNMi - Health-Related Fake News Mitigation project, selected in the NGI Search.

References

[1] C. Raffel, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research (2020).

[2] S. Minaee, T. Mikolov, N. Nikzad, M. A. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, ArXiv abs/2402.06196 (2024). URL: https://api.semanticscholar.org/CorpusID:267617032.

[3] A. Vaswani, et al., Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[4] OpenAI, Gpt-3: Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2021).

[5] E. M. Bender, et al., On the dangers of stochastic parrots: Can language models be too big?, FAccT (2021).

[6] A. Bruno, P. L. Mazzeo, A. Chetouani, M. Tliba, M. A. Kerkouri, Insights into classifying and mitigating llms' hallucinations, 2023. URL: https://arxiv.org/abs/2311.08117. arXiv:2311.08117.

[7] A. Bruno, et al., Insights into classifying and mitigating llms' hallucinations, in: CEUR WORKSHOP PROCEEDINGS, CEUR-WS, 2023.

[8] Z. Ji, et al., Hallucinations in large language models: Survey and taxonomy, arXiv preprint arXiv:2302.06453 (2023).

[9] J. Kalyanam, et al., Health misinformation detection in social media: A scalable system and lessons learned, in: WWW, 2019.

[10] D. A. Broniatowski, A. M. Jamison, S. C. Quinn, The misinformation pandemic: Covid-19's role in amplifying misinformation online, American Journal of Public Health 111 (2021) S235–S238. doi:10.2105/AJPH.2021.306466.

[11] X. Zhang, et al., Health misinformation on social media: A systematic review, Digital Health (2021).

[12] K. Shu, et al., Beyond news contents: The role of social context for fake news detection, in: WSDM, 2019.

[13] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, CoRR abs/2005.11401 (2020). URL: https://arxiv.org/abs/2005.11401. arXiv:2005.11401.

[14] P. Lewis, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, NeurIPS (2020).

[15] J. Lin, et al., Bertserini: A benchmarking toolkit for dense retrieval and ranking using bert, arXiv preprint arXiv:2101.12003 (2021).
[16] J. W. Rae, et al., Scaling language models: Methods, analysis, and insights from training gopher, arXiv preprint arXiv:2112.11446 (2021).
[17] I. Beltagy, et al., Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).
[18] T. B. Brown, et al., Language models are few-shot learners, NeurIPS (2020).
[19] R. S. Dudhabaware, M. S. Madankar, Review on natural language processing tasks for text documents, in: 2014 IEEE International Conference on Computational Intelligence and Computing Research, IEEE, 2014, pp. 1–5.
[20] R. Collobert, et al., Natural language processing (almost) from scratch, Journal of machine learning research 12 (2011) 2493–2537.
[21] K. Dong, et al., Multi-view content-aware indexing for long document retrieval, arXiv preprint arXiv:2404.15103 (2024).
[22] A. Jimeno Yepes, et al., Financial report chunking for effective retrieval augmented generation, arXiv preprint arXiv:2402.05131 (2024). URL: https://arxiv.org/abs/2402.05131. doi:10.48550/arXiv.2402.05131.
[23] Y. Zhang, Retrieval-augmented generation solutions for typical application process issues, in: Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024), SCITEPRESS – Science and Technology Publications, Lda, 2024, pp. 164–167. doi:10.5220/0012917400004508, paper published under CC BY-NC-ND 4.0 license.
[24] P. Biswas, et al., Document layout analysis in llms, arXiv preprint arXiv:2110.10177 (2021).
[25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.
[26] A. D. et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.