RE-FIN: Retrieval-based Enrichment for Financial data

Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica and Filippo Pallucchini
Dept of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy
CRISP Research Centre, University of Milano-Bicocca, Italy
{name.surname}@unimib.it

Abstract

Enriching sentences with knowledge from qualitative sources benefits various NLP tasks and enhances the use of labelled data in model training. This is crucial for Financial Sentiment Analysis (FSA), where texts are often brief and contain implied information. We introduce RE-FIN (Retrieval-based Enrichment for FINancial data), an automated system designed to retrieve information from a knowledge base to enrich financial sentences, making them more knowledge-dense and explicit. RE-FIN generates propositions from the knowledge base and employs Retrieval-Augmented Generation (RAG) to augment the original text with relevant information. A large language model (LLM) rewrites the original sentence, incorporating this data. Since the LLM does not create new content, the risk of hallucinations is significantly reduced. The LLM generates multiple new sentences using different relevant information from the knowledge base; we developed an algorithm to select one that best preserves the meaning of the original sentence while avoiding excessive syntactic similarity. Results show that enhanced sentences present lower perplexity than the original ones and improve performances on FSA.

1 Introduction

Financial sentiment analysis (FSA) aims to determine the sentiment conveyed in financial texts regarding a specific stock or the overall market outlook. To address the challenge posed by the market's active shifts, automated FSA has gained increasing attention in the past years (Van de Kauter et al., 2015). It has proven to be a powerful tool to support business decision-making and perform financial forecasting (Ma et al., 2023, 2024).

Nevertheless, FSA presents unique challenges with respect to general SA. The language in finance is highly specialized, filled with acronyms, technical jargon, industry-specific terms, and sarcasm, making it tricky for models to understand (Cambria et al., 2017; Malandri et al., 2018). Moreover, there's a shortage of large, labelled datasets, and annotating financial text requires expertise that's not easily scalable. Therefore, classification models often perform much worse in FSA than they do with more general SA (Xing et al., 2020). Even embedding alignment, which has proven effective in adapting models to specialized domains (D'Amico et al., 2024; Malandri et al., 2024) in certain fields, in FSA remains inconsistent (Liu et al., 2019).

Recently, Large Language Models (LLMs) have emerged as a potential tool to address the challenges mentioned above, offering powerful capabilities that can be applied to the financial domain. With their recent widespread adoption, models like ChatGPT and GPT-4 have demonstrated impressive performance in various NLP tasks(Bang et al., 2023; Omar et al., 2023; Khoury et al., 2023), including FSA (Li et al., 2023). However, directly applying LLMs for FSA poses two notable challenges. Firstly, the discrepancy between the objective function used in LLMs' pre-training and the goal of predicting financial sentiment may result in LLMs' inability to consistently output labels for financial sentiment analysis as expected (Ouyang et al., 2022; Thoppilan et al., 2022). Secondly, the typical subjects of financial sentiment analysis, such as news flashes and tweets, are characteristically concise and often lack adequate background information (Zhang et al., 2023; Van de Kauter et al., 2015).

To address the challenges above, we present a retrieval-augmented LLM framework for FSA. This paper proposes a new method to retrieve information from credible and customizable unstructured knowledge to enrich sentences. This approach makes the data more rich and understandable, which can increment user engagement—an essential factor in Machine Learning application (Cesarini et al., 2024) — and improves FSA. We

Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, pages 751–759
January 19–24, 2025. ©2025 Association for Computational Linguistics

can summarize our contributions in the following points:

- We present RE-FIN, a methodology for RAG that extracts propositions from a knowledge base and integrates them with original texts using LLMs through an innovative post-retrieval approach.
- Through evaluation on state-of-the-art benchmarks and an ablation study, we demonstrate that RE-FIN outperforms existing approaches.
- We provide the code freely to the community, promoting accessibility and further research[1] .

2 Related Works

2.1 FSA Models

Financial Sentiment Analysis (FSA) evaluates market sentiment by analyzing news and social media data, which can predict investment behaviors and equity market trends (Mishev et al., 2020). Understanding the effectiveness of these models in finance significantly impacts downstream financial analysis tasks (Li et al., 2023). Like other finance areas, such as named entity recognition and question-answering systems, LLMs are increasingly adopted in FSA (Li et al., 2023), enhancing the extraction of insights from unstructured data and improving decision-making. Early approaches (Araci, 2019; Day and Lee, 2016; Sohangir et al., 2018; Yang et al., 2020) utilized fine-tuned models achieving high performance but suffered from limited generalization due to reliance on specific training datasets (Xing, 2024). This highlights the need for more flexible models in FSA. Recent studies indicate that LLMs can outperform fine-tuned models in certain tasks. While these models exhibit strong generalization abilities as problem solvers (Li et al., 2023), applying them to FSA presents challenges (Zhang et al., 2023). Financial domain LLMs, such as BloombergGPT (Wu et al., 2023) and FinGPT (Yang et al., 2023), struggle to generate accurate sentiment labels due to a mismatch between their training objectives, typically Causal Language Modeling, and those of financial sentiment analysis (Zhang et al., 2023). Furthermore, financial sentiment analysis often addresses brief subjects like news flashes and tweets, which lack sufficient context, complicating reliable sentiment assessment. Implicit sentiment, where factual information suggests positive or negative sentiment, further complicates the issue (Van de Kauter et al., 2015).

2.2 RAG Models

LLMs demonstrate remarkable capabilities but face challenges such as hallucination, outdated knowledge, and opaque reasoning processes. Retrieval-Augmented Generation (RAG) provides a promising solution by incorporating knowledge from external databases, enhancing generation accuracy and credibility, particularly for knowledge-intensive tasks, while enabling continuous updates and integration of domain-specific information (Gao et al., 2023). RAG (Cai et al., 2022; Lewis et al., 2020) merges the strengths of context retrieval and LLMs for language generation (Zhang et al., 2023). This method leverages two distinct knowledge sources: the parametric memory within LLMs and the nonparametric memory from retrieved documents, effectively guiding generation to yield more accurate, contextually relevant responses. RAG has seen extensive application in open-world QA (Mao et al., 2021) and code summarization (LIU et al., 2021; Parvez et al., 2021).

The success of RAG heavily depends on the quality of the retrieval process, which employs sentence embeddings (Salemi and Zamani, 2024). While sentence embeddings capture overall text meaning as fixed-length representations (Morris et al., 2023), querying them for semantic information at a granular level is challenging (Rudinger et al., 2017; Qin and Van Durme, 2023; Wang and Yu, 2023). This limitation restricts expressivity in tasks like document retrieval, particularly when identifying concepts expressed in specific document segments rather than the entire document. Previous studies have shown success with phrase retrieval or late-interaction models that provide more granular representations of the retrieval corpus (Seo et al., 2019; Khattab and Zaharia, 2020; Lee et al., 2021a,b). Coarse-grained retrieval units may deliver relevant information, yet they risk introducing redundant content that could distract retrievers and LLMs in downstream tasks (Yu et al., 2023; Shi et al., 2023), especially in sentiment classification, where excessive information might confuse rather than clarify. Therefore, we adopt a fine-grained retrieval logic utilizing document propositions, computed using the model developed by Chen et al. (Chen et al., 2023). Propositions represent atomic expressions in the text, encapsulating unique factual segments in concise, self-contained natural language (Gao et al., 2023). Additionally, the generation process itself can pose challenges.

[1] https://github.com/filippopallucchini/RE-FIN

For example, concepts in external documents may be similar but not identical to those in the question, which could mislead the LLM (Chen et al., 2024). While most approaches focus on controlling the retrieval process, evaluating the generation is equally important (Cheng et al., 2024; Es et al., 2023). To address this, instead of generating a single augmented sentence, we produce multiple options and select the most suitable one using a post-retrieval method developed in this work.

3 Methods

Here, we describe the framework of the model proposed in this paper and sketched in Fig. 1. A traditional RAG process includes three main phases indexing, retrieval, and generation; moreover, an advanced RAG method, like the one proposed in the paper, also employs pre-retrieval and post-retrieval strategies (Gao et al., 2023).

Indexing starts with the cleaning and extraction of raw data in PDF, CSV, and TSV formats and converts them into a uniform plain text format. To accommodate the context limitations of language models, text is segmented into sentences delimited by points, becoming smaller and digestible chunks.

Pre-retrieval process. In this stage, the primary focus is optimizing the indexing structure and the original query. Optimizing indexing aims to enhance the quality of the content being indexed. We involve a very little-used strategy proposed by Chen et al. (Chen et al., 2023), enhancing data granularity, and optimizing index structures. We choose propositions as a retrieval unit since the retrieved texts are more condensed with information relevant to the original sentence, reducing the need for lengthy input tokens and minimizing the inclusion of extraneous, irrelevant information. Propositions are then encoded into vector representations using an embedding model and stored in a vector database. This step enables efficient similarity searches in the subsequent retrieval phase.

Retrieval. Upon receipt of a user query, the RAG system employs the same encoding model utilized during the indexing phase to transform the query into a vector representation. It then computes the similarity scores between the query vector and the vector of chunks within the indexed corpus. The system prioritizes and retrieves the top K chunks that demonstrate the greatest similarity to the query. These chunks are subsequently used as the expanded context in the prompt.

Post-Retrieval Process. Once the relevant context is retrieved, it's crucial to integrate it effectively with the query. The main methods in the post-retrieval process include re-ranking chunks and context compressing. In particular, we utilize an innovative heuristic process to create a new sentence similar to the original one that includes the most relevant documents retrieved.

Generation. In this phase, the best sentence enriched created is corrected using an LLM and used as the final version of the sentence.

Now, we are going to describe the method more analytically. Let's consider I as the set of original sentences to be enriched, where $i \in I$, and K as the set of sentences from the knowledge corpus chosen for enriching the original sentences, where $k \in K$. We choose knowledge data from Investopedia downloaded from huggingface platform[2][3], from two of the most important book of Finance (Fisher, 2003; Graham and McGowan, 2005) and the dataset of financial terms definitions provided by Ghosh et al. (Ghosh et al., 2022). The first task is to extract the propositions that compose the sentences of both the original text and the knowledge. To perform this we utilize the Propositionizer proposed by Chen et al, 24 (Chen et al., 2023)[4], that we call $PROP$, such that

$$p_i = PROP(i) \tag{1}$$

where $p_i = (1, ..., n^i)$ and

$$p_k = PROP(k) \tag{2}$$

where $p_k = (1, ..., n^k)$. Now we use these propositions to retrieve for each i the most similar document from the knowledge, exploiting the $Cosine\ Similarity\ CS_{p_i p_k} = cosim(E(p_i), E(p_k)))$ such that:

$$r_i = \max_{k=1}^{K}(CS_{p_i p_k} | CS_{p_i p_k} > \beta) \tag{3}$$

where r_i is the set of documents retrieved and E is encoder-only model[5] provided by huggingface. β is a constraint designed to retrieve just those documents composed by a proposition semantically very similar to one proposition of the original text. We add two other constraints to take under control that:

[2] https://huggingface.co/datasets/infCapital/investopedia_terms_en

[3] https://huggingface.co/datasets/openvega-simon/investopedia

[4] https://huggingface.co/chentong00/propositionizer-wiki-flan-t5-large

[5] https://huggingface.co/intfloat/e5-base-v2

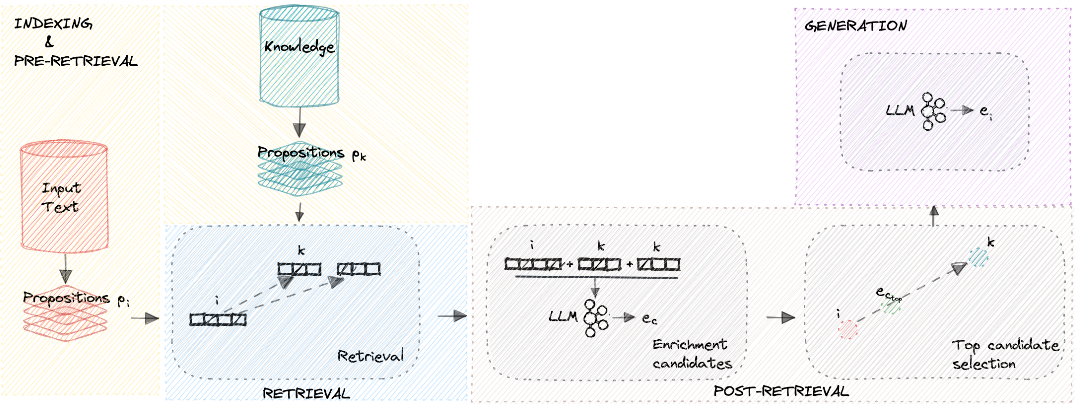

Figure 1: Diagram of the proposed model called RE-FIN.

- k would not be too similar to p_i (using γ); because, in this case, the retrieval could be useless

$$r_i = \max_{k=1}^{K}(CS_{kp_i} | CS_{kp_i} < \gamma) \tag{4}$$

- i would not be too different to p_k (using ϵ) and k itself (using $\epsilon * 1.2$)); because, in this case, the retrieval could be not adequate if not pejorative

$$r_i = \max_{k=1}^{K}(CS_{ip_k} | CS_{ip_k} > \epsilon) \tag{5}$$

$$r_i = \max_{k=1}^{K}(CS_{ik} | CS_{ik} > \epsilon * 1.2) \tag{6}$$

k document that respects all constraints indicated will be used for the enrichment task. This task is performed with the fundamental aid of a decoder-only model, that comes from the work of Jiang et al. (Jiang et al., 2023) and it is provided by HuggingFace[6], that merges i and k to create the enriched sentence e. e is the result of three steps:

- produce ζ candidates

  $$e_c = LLM(i, r_i) \tag{7}$$

  where $c = (1, ..., \zeta)$
- select $zeta_{top}$ candidates closest to a reference vector v_i positioned between $E(i)$ and $E(r_i)$ with a μ pace calculated with a *Move Towards* (MT) function

  $$v_i = MT(E(i), E(r_i) | \mu) \tag{8}$$

  such that

  $$e_{ctop} = \max_{\zeta_{top}}(CS_{e_c v_i}) \tag{9}$$
- correct ζ_{top} using the Eq.7 with prompt adjusted and select e with the Eq.9 respecting two last constraints

  $$e_i = \max_{1}^{\zeta_{top}}(CS_{e_{c_{top}} v_i} | CS_{ie_i} > \omega) \tag{10}$$

where ω is the minimum semantic similarity between $E(i)$ and $E(e_i)$. The system described above allows us to use propositions for precise retrieval while utilizing the entire document for enrichment. This process is enabled by a controlled step that verifies: (i) the semantic similarity of the entire document relative to the original sentence (as described in Eq.6), and (ii) the semantic similarity of the enriched sentence in relation to the original sentence (as described in Eq.10).

4 Evaluation

For the evaluation, we selected three datasets well highly used for FSA, as in the papers we used as main references (Xing, 2024; Li et al., 2023; Zhang et al., 2023; Du et al., 2024). **Financial PhraseBank (FPB)** (Malo et al., 2014). FPB includes 4,846 news annotated by 16 individuals with adequate background knowledge of financial markets from an investor perspective. Based on the strength of agreement among annotators, it releases four reference datasets, namely 100%, 75%, 66%, and 50% agreement. In their study, Malo et al. argues that the overall sentiment may be different from the prior sentiment polarity of individual words, and incorporating phrase-structure information and domain-specific use of language could improve the detection. We use the 100% agreement dataset. **FiQA Task 1** (Maia et al., 2018). The dataset is from FiQA Open Challenge Task 1, which consists of 498 financial news headlines and 675 posts

[6] https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GPTQ

Dataset	FPB	FiQA	SEntFiN
Positive	570	507	2832
Negative	303	264	2373
Neutral	1391	-	2701
Total Size	2264	771	7906

Table 1: Summary statistics for the three FSA datasets (after post-processing).

with their target entities, aspects, and corresponding sentiment score. The original dataset has 1173 messages with sentiment scores ranging from -1 to +1. By filtering those scores with an absolute value larger than 0.3, only 771 messages are left and mapped to the positive/negative classes exactly as (Xing, 2024). **SEntFiN 1.0** (Sinha et al., 2022). SEntFiN is a human-annotated dataset that includes 10,753 news headlines with their entity and corresponding sentiment. Commonly, multiple entities are present in a news headline with different sentiment expressions and SEntFiN has 2,847 headlines that contain multiple entities, which may have conflicting sentiment. For this reason, we consider in our experiment just those documents without conflict.

We conduct our evaluation over 3 different tasks without and with the enrichment process:
FSA with decoder-only: Predict sentiment of sentences through a pre-trained LLM.
FSA with encoder-only: Fine-tune and predict sentiment through a pre-trained encoder-only model
Perplexity
The parameters utilized for the experiments were deducted from a sensitivity analysis. We select these values as optimal: $\beta = 0.8$, $\gamma = 0.95$, $\epsilon = 0.7$, $\zeta = 50$, $\mu = 0.12$, $\zeta_{top} = 5$, $\omega = 0.83$.

4.1 FSA

First, we tested whether a decoder-only model is better at predicting the sentiment of a sentence after enrichment. We prompted the input sentence, asking the LLM[6] to predict the sentiment as either (POSITIVE, NEGATIVE) or (POSITIVE, NEUTRAL, NEGATIVE), depending on the dataset used, employing both zero-shot and few-shot learning. For the few-shot scenario, we randomly selected one example per label from the respective dataset: 3 examples for FPB, 6 for SEntFiN (as it contains twice as many as FPB), and 2 for FIQA. Specifically, we compared the model's accuracy in predicting the sentiment of the dataset with and without enriched sentences to assess whether enrichment aids a pre-trained model in predicting a sentence's financial sentiment. Following this, we conducted another FSA using a pre-trained encoder-only model[7] that was fine-tuned without and with the enriched sentences.

Dataset	FPB	FiQA	SEntFiN
Decoder-only - Zero-shot			
Mistral	75.8%	79.1%	65.4%
Mistral + RE-FIN	86.4%	**87.3%**	68.1%
Decoder-only - Few-shot			
Mistral	87.6%	79.9%	66.9%
Mistral + RE-FIN	91.5%	**87.3%**	69.6%
Encoder-only - Fine-tuning			
DistilBert	90.6%	73.1%	58.8%
DistilBert + RE-FIN	**92.8%**	73.1%	**71.9%**

Table 2: Accuracy for FSA using the encoder-only model, considering only the enriched documents for each dataset.

Dataset	FPB	FiQA	SEntFiN
Decoder-only - Zero-shot			
Mistral	75.1%	80.9%	70.7%
Mistral + RE-FIN	79.3%	83.9%	71.1%
Decoder-only - Few-shot			
Mistral	86.3%	80.3%	67.7%
Mistral + RE-FIN	88.0%	82.4%	68.0%
Encoder-only - Fine-tuning			
DistilBert	93.2%	71.0%	86.0%
DistilBert + RE-FIN	**95.8%**	**85.4%**	**86.7%**

Table 3: Accuracy for FSA. The accuracy reported for the Encoder-only evaluation was computed after 1 epoch.

It is easier to notice the performance increase due to RE-FIN. On average there is an increase of 3.8% utilizing a zero-shot prompt and 4.3% with few-shot learning. The sole exception is the encoder-only model trained exclusively on the augmented data of the FiQA dataset, which exhibits the same performances achieved with the non-augmented data. Nonetheless, decoder-only models that employ data augmented via RE-FIN achieve the highest overall performance for this dataset. Both encoder-only and decoder-only models were chosen with the belief that they offer a strong foundation while being accessible and easy to use for the entire community.

4.2 Perplexity

Perplexity is a measurement that reflects how well a model can predict the next word based on the preceding context. So, we thought that computing the perplexity w/o enrichment could give a reliable

[7] https://huggingface.co/distilbert/distilbert-base-uncased

measure of improving the clarity and objective of sentences. We utilized a commonly used Python library *evalute*[8], testing the two most used LLMs provided: *openai-gpt* and *gpt2*.

Dataset	FPB	FiQA	SEntFiN
openai-gpt	531.4	4572.7	6072.7
openai-gpt + RE-FIN	384.3	3099.3	5427.3
gpt2	180.1	1162.5	1219.4
gpt2 + RE-FIN	**138.2**	**670.5**	**1090.2**

Table 4: Mean Perplexity.

4.3 Ablation Analysis

We conducted an ablation study to evaluate the robustness of our model and the contribution of each component. We tested each dataset with three distinct settings to demonstrate the value of each component of our method. The experiments were carried out on the decoder-only fine-tuning task, as it yielded the best performance, as discussed in the previous section. The approaches we tested are:

- No Retrieval: Sentences are enriched directly using the LLM[6], without any retrieval process or additional steps.
- No Post-Retrieval: The retrieval process is applied as described in Sec.3, but sentences are enriched with the LLM[6] without the post-retrieval phase.
- No MT: The complete method is used, except the MT logic, which is responsible for selecting the best-enriched candidate, is removed. Instead, a simpler function based on cosine similarity is used to select the candidate most similar to the original sentence.

Fig.2 shows the contribution of the retrieval process compared to enriching sentences without it. The most notable insight from the results is the significant impact of candidate selection criteria. Relying solely on cosine similarity resulted in the lowest accuracy for two out of three datasets, emphasizing the importance of our MT function in selecting the best-enriched candidate.

5 Results

RE-FIN demonstrates superior performance across all datasets and classification methods tested, with the only exception being the encoder-only model trained with augmented data from the FiQA dataset, which performs similarly to non-augmented data.

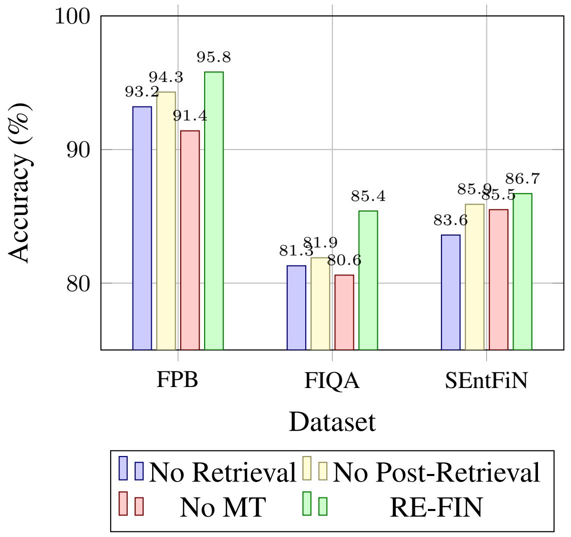

Figure 2: Accuracy across datasets (FPB, FIQA, SEntFiN) for different iterations.

However, the highest performance for this dataset is achieved by decoder-only models utilizing RE-FIN augmented data. The ablation study in Fig.2 shows that every component of RE-FIN positively contributes to overall performance, emphasizing its effectiveness in enhancing classification results. Notably, RAG and fine-tuning are not mutually exclusive but can complement each other, enhancing models at different levels (Gao et al., 2023). For FPB and SEntFin, their combined use achieves optimal performance.

6 Conclusions

In this paper, we developed a novel RAG methodology that enriches domain-specific sentences with reliable, knowledge-based information. Our model retrieves information based on propositions, seeking sentences that share similar propositions while providing added value. Additionally, it introduces an novel selection criterion to choose the candidate that best integrates the input sentence with information from retrieved documents. Experimental results on three FSA datasets show that RE-FIN consistently improves sentiment analysis performance across all datasets, achieving superior accuracy compared to existing methods. The ablation study indicates that each component of RE-FIN enhances its overall effectiveness. The RE-FIN tool is released as a free and open-source resource for the research community[9], enabling broader access and advancing financial sentiment analysis.

[8] https://pypi.org/project/evaluate/

[9] https://github.com/filippopallucchini/RE-FIN

Acknowledgments

This work is partially supported within the research activity of an Italian Project entitled "ISALDI: Interpretable Stock Analysis Leveraging Deep multImodal models" - PRIN 2022 Project number 2022P4MPAP - in which some of the authors are involved as coordinators and researchers.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718.

Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. 2022. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 3417–3419.

Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Mirko Cesarini, Lorenzo Malandri, Filippo Pallucchini, Andrea Seveso, and Frank Xing. 2024. Explainable ai for text classification: Lessons from a comprehensive evaluation of post hoc methods. *Cognitive Computation*, pages 1–19.

Sihao Chen, Hongming Zhang, Tong Chen, Ben Zhou, Wenhao Yu, Dian Yu, Baolin Peng, Hongwei Wang, Dan Roth, and Dong Yu. 2024. Sub-sentence encoder: Contrastive learning of propositional semantic representations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1596–1609.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2023. Dense x retrieval: What retrieval granularity should we use? *arXiv preprint arXiv:2312.06648*.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2024. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36.

Simone D'Amico, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Filippo Pallucchini. 2024. Alignment of multilingual embeddings to estimate job similarities in online labour market. In *2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Min-Yuh Day and Chia-Chou Lee. 2016. Deep learning for financial sentiment analysis on finance news providers. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1127–1134. IEEE.

Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Philip A Fisher. 2003. *Common stocks and uncommon profits and other writings*, volume 40. John Wiley & Sons.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Sohom Ghosh, Shovon Sengupta, Sudip Kumar Naskar, and Sunny Kumar Singh. 2022. Finrad: Financial readability assessment dataset-13,000+ definitions of financial terms for measuring readability. In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pages 1–9.

Benjamin Graham and Bill McGowan. 2005. *The intelligent investor*. Harper Collins New York.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Raphaël Khoury, Anderson R Avila, Jacob Brunelle, and Baba Mamadou Camara. 2023. How secure is code generated by chatgpt? In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2445–2451. IEEE.

Jaewoong Lee, Heejoon Lee, Hwanhee Lee, and Kyomin Jung. 2021a. Learning to select question-relevant relations for visual question answering. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 87–96.

Jinhyuk Lee, Alexander Wettig, and Danqi Chen. 2021b. Phrase retrieval learns passage retrieval, too. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3661–3672.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422.

Ruijun Liu, Yuqian Shi, Changjiang Ji, and Ming Jia. 2019. A survey of sentiment analysis based on transfer learning. *IEEE access*, 7:85401–85412.

Shangqing LIU, Yu CHEN, Xiaofei XIE, Jingkai SIOW, and Yang LIU. 2021. Retrieval-augmented generation for code summarization via hybrid gnn.(2021). In *Proceedings of the Ninth International Conference on Learning Representations: ICLR*, pages 4–8.

Yu Ma, Rui Mao, Qika Lin, Peng Wu, and Erik Cambria. 2023. Multi-source aggregated classification for stock price movement prediction. *Information Fusion*, 91:515–528.

Yu Ma, Rui Mao, Qika Lin, Peng Wu, and Erik Cambria. 2024. Quantitative stock portfolio optimization by multi-task learning risk and return. *Information Fusion*, 104:102165.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Filippo Pallucchini. 2024. Sense: embedding alignment via semantic anchors selection. *International Journal of Data Science and Analytics*, pages 1–15.

Lorenzo Malandri, Frank Z Xing, Carlotta Orsenigo, Carlo Vercellis, and Erik Cambria. 2018. Public mood–driven asset allocation: The importance of financial sentiment in portfolio management. *Cognitive Computation*, 10(6):1167–1176.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100.

Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir T Chitkushev, and Dimitar Trajanov. 2020. Evaluation of sentiment analysis in finance: from lexicons to transformers. *IEEE access*, 8:131662–131682.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460.

Reham Omar, Omij Mangukiya, Panos Kalnis, and Essam Mansour. 2023. Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. *arXiv preprint arXiv:2302.06466*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Md Rizwan Parvez, Wasi Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2719–2734.

Guanghui Qin and Benjamin Van Durme. 2023. Nugget: Neural agglomerative embeddings of text. In *International Conference on Machine Learning*, pages 28337–28350. PMLR.

Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Skip-prop: Representing sentences with one vector per proposition. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)—Short papers*.

Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2395–2400.

Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. Real-time open-domain question answering with dense-sparse phrase index. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4430–4441.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Ankur Sinha, Satishwar Kedas, Rishu Kumar, and Pekka Malo. 2022. Sentfin 1.0: Entity-aware sentiment analysis for financial news. *Journal of the Association for Information Science and Technology*, 73(9):1314–1335.

Sahar Sohangir, Dingding Wang, Anna Pomeranets, and Taghi M Khoshgoftaar. 2018. Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5(1):1–25.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Marjan Van de Kauter, Diane Breesch, and Véronique Hoste. 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with applications*, 42(11):4999–5010.

Hongwei Wang and Dong Yu. 2023. Going beyond sentence embeddings: A token-level matching algorithm for calculating semantic textual similarity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 563–570.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Frank Xing. 2024. Designing heterogeneous llm agents for financial sentiment analysis. *arXiv preprint arXiv:2401.05799*.

Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: an investigation into common mistakes and silver bullets. In *Proceedings of the 28th international conference on computational linguistics*, pages 978–987.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *FinLLM at IJCAI*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 349–356.