Self-explanatory and Retrieval-augmented LLMs for Financial Sentiment Analysis

Filippo Pallucchini
University of Milano-Bicocca
Italy
filippo.pallucchini@unimib.it

Xulang Zhang
Nanyang Technological University
Singapore
xulang.zhang@ntu.edu.sg

Rui Mao
Nanyang Technological University
Singapore
rui.mao@ntu.edu.sg

Erik Cambria
Nanyang Technological University
Singapore
cambria@ntu.edu.sg

Abstract

Enriching sentences with qualitative knowledge is crucial for enhancing sentiment prediction and making the most of the available labelled data for training models. This is particularly important in domains like the financial one, where texts are usually brief and contain much-implied information. In this article, we introduce FLEX (Financial Language Enhancement with Guided LLM Execution), an automated system capable of retrieving information from a Large Language Model (LLM) to enrich financial sentences, making them more knowledge-dense and explicit. FLEX generates multiple potentially enhanced sentences and uses a new logic to determine the most suitable one. To mitigate hallucinations in LLMs, we developed a new algorithm to select the most appropriate sentences. This approach ensures the original meaning is preserved, reduces excessive syntactic similarity between versions, and maintains the lowest possible perplexity. These enhanced sentences are more interpretable and directly useful for downstream tasks like financial sentiment analysis (FSA). Compared to state-of-the-art methods, FLEX shows improvements in the accuracy of processing FSA tasks.

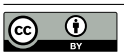
CCS Concepts

• **Information systems** → **Information retrieval**; **Sentiment analysis**; • **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → **Natural language generation**.

Keywords

Financial sentiment analysis, LLM, RAG, semantics, interpretability

ACM Reference Format:
Filippo Pallucchini, Xulang Zhang, Rui Mao, and Erik Cambria. 2025. Self-explanatory and Retrieval-augmented LLMs for Financial Sentiment Analysis. In *The 40th ACM/SIGAPP Symposium on Applied Computing (SAC '25), March 31-April 4, 2025, Catania, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3672608.3707894



SAC '25, March 31-April 4, 2025, Catania, Italy

ACM ISBN 979-8-4007-0629-5/25/03
https://doi.org/10.1145/3672608.3707894

1 Introduction

Financial sentiment analysis (FSA), which broadly encompasses the study of investor sentiment and financial textual sentiment [12], is a key domain within sentiment analysis. Given the complex nature of financial markets, individuals involved in different market conditions display varied cognitive patterns [28, 30], making it difficult to dynamically understand and analyze the market for sound financial decision-making. To tackle the challenges arising from the market's constant fluctuations, automated FSA has garnered significant attention over the past decade [47]. It has proven to be a powerful tool for supporting business decision-making and conducting financial forecasting [23, 24]. Applications include corporate disclosures, annual reports, earnings calls, financial news, social media interactions, and more [47, 51]. Sentiment analysis is a complex, domain-dependent problem. This domain dependence is especially pronounced in the finance sector [27] due to the focused nature of financial topics and the use of highly specialized language [32, 34, 59]. For instance, words like *liability* and *debt* are typically viewed negatively in general-purpose sentiment analysis, but they often carry a neutral meaning in the financial context [12, 52]. Some authors have tackled these challenges through embedding alignment, which has proven effective in adapting models to specialized domains [8, 26], although performance remains predominantly inconsistent in the field of FSA [21]. Large Language Models (LLMs) have matured significantly and gained widespread adoption across various domains and everyday tasks [42]. This advancement has also had a profound impact on the financial industry.

Recent models, such as ChatGPT[1] and GPT-4[2], trained with reinforcement learning from human feedback (RLHF) [7] and masked language model objectives, have demonstrated exceptional capabilities across a wide range of natural language processing (NLP) tasks[2, 29, 39]. These LLMs are trained on datasets covering diverse genres and topics. While their performance in general NLP tasks is impressive, their applicability and effectiveness in specific domains like finance still require further exploration, as their impact could span a wide range of applications [11, 19]. In the financial domain, LLMs are increasingly crucial for tasks such as investment sentiment analysis, financial named entity recognition, and question-answering systems to assist financial analysts [19].

[1] https://platform.openai.com/docs/models/gpt-3-5
[2] https://platform.openai.com/docs/models/gpt-4

However, directly applying LLMs for FSA presents two main challenges. First, the difference between the objective functions used in LLMs' pre-training and the goal of predicting financial sentiment can cause LLMs to inconsistently output labels for FSA [40, 46]. Second, the typical sources for FSA, like news flashes and tweets, are often concise and lack sufficient background information [47, 56]. This information scarcity not only affects human judgment [27] but also poses a significant challenge for LLMs in making accurate predictions [56].

To tackle these challenges, our study introduces a retrieval-augmented LLM framework for financial sentiment analysis. Similar to how a makeup artist enhances a person's features [61], we propose a novel method for making a sentence more understandable and self-explanatory without altering its essence, which is critical for real-world AI applications [4, 5]. This approach highlights financial concepts and implicit propositions by retrieving relevant information from an LLM, thereby improving FSA. We provide experimental evidence, demonstrating the model's effectiveness on two analytical tasks, namely, perplexity and FSA, using three benchmark datasets.

The contributions of this work can be summarized as follows:

- We propose a novel approach that incorporates semantic similarity and perplexity to enhance the decision-making logic and interpretability of a predictive model for FSA tasks.
- We demonstrate that our model's enrichment approach improves results, even when using the original dataset for fine-tuning models or making direct predictions.
- We provide the code freely to the community, promoting accessibility and further research[3]

2 Related Works

2.1 FSA Models

Sentiment analysis is one of the most commonly used NLP techniques in the financial sector, often employed to predict investment behaviors and trends in equity markets based on news and social media data [6, 37]. Understanding the effectiveness of such models in the financial domain could significantly impact many downstream financial analytical tasks [19]. LLMs are increasingly playing a crucial role in financial tasks such as investment sentiment analysis, financial named entity recognition, and question-answering systems that assist financial analysts [19]. Early approaches [1, 9, 45, 54] involved fine-tuning models, some of which achieved high performance. However, these models are often too dependent on the specific datasets used for fine-tuning (evidence shows that FinBERT performs poorly on certain datasets), highlighting the need for more flexible, ready-to-use models for FSA. Recent studies indicate that some LLMs outperform fine-tuned models on certain tasks. While these LLMs exhibit unprecedented quality, retain accumulated knowledge, and demonstrate excellent generalization ability as general problem solvers [19], their effectiveness in financial sentiment analysis remains critically important. Biased predictions can also be associated with language models due to the uneven label distribution [35], sense ambiguity [60], and language preference [58] for sentiment analysis tasks.

[3]https://github.com/filippopallucchini/FLEX

LLMs specifically designed for the financial domain, such as BloombergGPT [49] and FinGPT [53], struggle to generate the expected sentiment labels due to a mismatch between their training objectives, typically Causal Language Modeling, and the specific requirements of financial sentiment analysis [56]. Furthermore, financial sentiment analysis often involves brief content, such as news flashes and tweets, which typically lack sufficient background information. This brevity and lack of context present a significant challenge for LLMs, complicating reliable sentiment analysis. Additionally, a well-known challenge of LLMs is their inconsistent reliability in the information they provide. Since financial texts frequently contain implicit sentiment—where factual information implies a positive or negative sentiment—it is crucial to address these challenges to improve the accuracy of FSA [47].

2.2 RAG Models

LLMs showcase impressive capabilities but encounter challenges like hallucination, outdated knowledge, and non-transparent, untraceable reasoning processes. Retrieval-Augmented Generation (RAG) has emerged as a promising solution by incorporating knowledge from external databases. This enhances the accuracy and credibility of the generation, particularly for knowledge-intensive tasks, and allows for continuous knowledge updates and integration of domain-specific information [13]. RAG [3, 18] is a technique that combines the strength of context retrieval and LLMs for language generation [56]. The combination of retrieval and generation in RAG allows it to utilize two distinct sources of knowledge: the parametric memory stored in the LLMs' parameters and the nonparametric memory obtained from the corpus of retrieved documents. This dual-knowledge approach enables RAG to effectively guide the generation process and produce more accurate and contextually relevant responses. RAG has been widely used in areas such as open-world QA [20, 36] and code summarization [22, 41]. Although sentence embeddings usually capture the meaning of the entire text sequence as a fixed-length embedding [33, 38], it is difficult in practice to query sentence embeddings for semantic information or structure on a more granular level [43, 48]. The format would offer limited expressivity when modeling tasks such as document retrieval, especially when the task conceptually involves identifying the document parts that respond to the query. For such reason, previous studies have found empirical success with phrase retrieval or late-interaction models, which support more granular and expressive representations of the retrieval corpus [15–17]. Similarly, approaches like Generate-Read [55] replace traditional retrieval with LLM-generated content, finding that LLM-generated contexts often contain more accurate answers due to better alignment with the pre-training objectives of causal language modeling. They use a prediction module that aims to reduce redundancy and noise by generating context directly through the LLM [55].

3 Method

The framework of our proposed model is sketched in Fig. 1, which consists of two main phases. 1) The **MAKEUP phase** generates enriched sentence candidates that maintain the exact semantics of the original text while clarifying specific financial concepts or making implicit propositions more explicit.

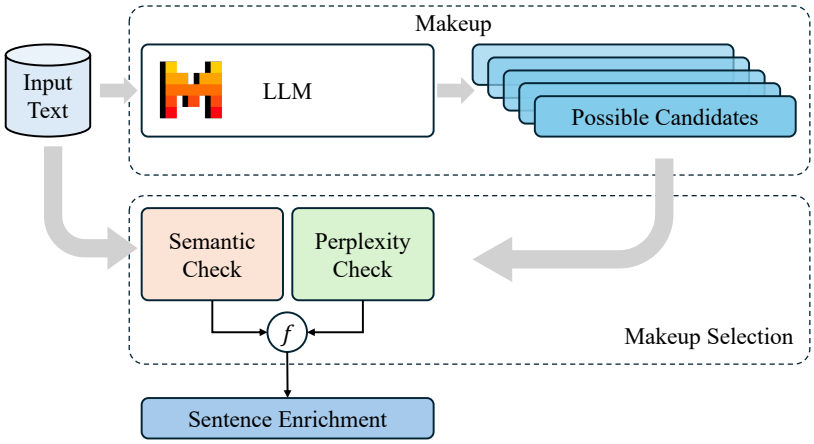


Figure 1: Diagram of the proposed FLEX method.

2) The **MAKEUP SELECTION phase** consists of a function that selects the most appropriate candidate from those produced in the previous phase. Using an embedding model, the function encodes both the original sentence and each candidate into vector representations to assess semantic similarity. In addition to semantic similarity, we also consider it crucial to ensure the sentence is clear and self-explanatory. Thus, the function also incorporates a measure of perplexity, which reflects how well the sentence aligns with the language model's expectations, providing insight into its readability and naturalness.

Specifically, given I is the set of original sentences to be enriched, where $i \in I$, the sentence is passed to a decoder-only model, that comes from the work of Jiang et al. [14] and it is provided by HuggingFace[4]. It performs the MAKEUP phase that consists in creating ζ possible candidates enriched e

$$e_c = LLM(i), \tag{1}$$

where $c = (1, ..., \zeta)$. The process of MAKEUP SELECTION is performed by a function that considers both the semantic similarity and the perplexity. In the *Sematic Check*, each i and e_c is embedded in an encoder-only model[5] E, provided by huggingface, and used to compute cosine similarity CS (*Cosine Similarity*)

$$CS_{ie_c} = cosim(E(i), E(e_c))) \tag{2}$$

where $CS_{ie_c} \in [0, 1]$.

At the same time, in the *Perplexity Check*, each i and e_c are passed to the commonly used Python library *evalute*[6] parametrized with *openai-gpt*, producing P (*Perplexity*)

$$P_{e_c} = Perplexity(e_c) \tag{3}$$

where $P_{e_c} \in [0, \infty]$.

Thus, the MS function combines these two measures to identify the best candidates that both maximize the semantic similarity score CS_{ie_c} and minimize the perplexity P_{e_c}. The function is defined as:

$$e_i = \max_{c=1}^{\zeta} \left(CS_{ie_c} + \left(1 - \frac{\log(1+P_{e_c})}{\log(1+P_{e_{c_{\max}}})} \right) \middle| \, CS_{ie_c} > \omega, \; P_{e_c} < P_i \right) \tag{4}$$

where the condition $CS_{ie_c} > \omega$ ensures that the selected enriched sentences have a minimum threshold of semantic similarity to the original sentence. This ω value depends on the maximum level of enrichment that we want to allow. Indeed, if a Dataset is composed by short sentences we will let more contribute to the LLM and vice-versa. The second condition, $P_{e_c} < P_i$, guarantees that the chosen sentence is clearer than the original one. The system described above allows us to control the process of enrichment avoiding as much as possible the risk taken by the LLM.

4 Experiments

4.1 Datasets

Three datasets are used for the evaluation process. They are the most used dataset in FSA in literature, particularly in the papers we used as main references like [10, 19, 50, 56]. The statistics of the employed datasets can be viewed in Table 1.

Financial PhraseBank (FPB) [27]. In 2014, a milestone dataset, i.e., FPB, was established, which includes 4,846 news annotated by 16 individuals who have adequate background knowledge in 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, [27] 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. The average sentence length of the dataset is 128.1 so we set the ω value at 0.72, a quite high value that limits the wide contribution of LLM but allows enrichment to take place.

FiQA Task 1 [25]. The dataset is from FiQA Open Challenge Task 1, which consists of 498 financial news headlines and 675 posts 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

[4]https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GPTQ
[5]https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[6]https://pypi.org/project/evaluate/

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

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

larger than 0.3, only 771 messages are left and mapped to the positive/negative classes exactly as [50]. The average sentence length of the dataset is 74.3 so we set the ω value at 0.43 (proportionate concerning the value used for FPB), a quite low value that allows a big contribution of LLM.

SEntFiN 1.0 [44]. To address the problem of scant benchmark dataset for fine-grained FSA, a challenging task that requires extensive human efforts for annotation, [44] released SEntFiN 1.0 and made it publicly available to promote further research. 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 [44]. For this reason, we consider just those documents without conflict in our experiment. The average sentence length of the dataset is 57.1 so we set the ω value at 0.2 (proportionate concerning the value used for FPB), a quite low value that allows a big contribution of LLM.

4.2 Setups

We evaluate our approach using three different frameworks, all without the enrichment process:

- FSA with decoder-only zero-shot: Predict sentence sentiment using a pre-trained LLM in a zero-shot setting.
- FSA with decoder-only few-shot: Predict sentence sentiment using a pre-trained LLM with a few-shot prompt.
- FSA with encoder-only: Fine-tune and predict sentiment using a pre-trained encoder-only model.

First of all, we tested if a decoder-only model is facilitated in predicting the sentiment of a financial sentence after the enrichment. So, we prompt the input sentence asking the LLM[4] to predict the sentiment (POSITIVE, NEGATIVE) or (POSITIVE, NEUTRAL, NEGATIVE) depending on the nature of the dataset used both with zero-shot learning and few-shot learning.

For the few-shot scenario, we randomly selected one example per label from the respective dataset: 3 for FPB, 6 for SEntFiN (which 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 an encoder-only model that was pre-trained and fine-tuned without enriched sentences.

Table 2: Performance comparison measured by accuracy.

Dataset	FPB	FiQA	SEntFiN	Avg.
Decoder-only - Zero-shot				
Mistral	75.1%	80.9%	70.9%	75.7%
Mistral + FLEX v0.2	83.5%	86.8%	73.7%	81.3%
Mistral + FLEX v0.3	85.3%	89.5%	76.1%	83.6%
Decoder-only - Few-shot				
Mistral	87.2%	80.5%	67.7%	78.5%
Mistral + FLEX v0.2	90.1%	87.9%	75.3%	84.4%
Mistral + FLEX v0.3	90.3%	90.3%	77.3%	86.0%
Encoder-only - Fine-tuning				
DistilBert	93.2%	71%	86.0%	83.4%
DistilBert + FLEX v0.2	94.7%	82.6%	84.4%	87.2%
DistilBert + FLEX v0.3	95.4%	91.6%	86.0%	91.0%

4.3 Baselines

We examine our method by comparing to the following baselines:
Mistral 7B is a 7-billion-parameter language model optimized for high performance and efficiency in NLP tasks. It incorporates several technical innovations that enhance its functionality, including Grouped-Query Attention, Sliding Window Attention, and advanced fine-tuning capabilities. We examine two versions of Mistral, namely, Mistral-7B-Instruct-v0.2-GPTQ and Mistral-7B-Instruct-v0.3-GPTQ for generating the enrichment, respectively. These models are termed FLEX v0.2 and FLEX v0.3. Then, we use Mistral-7B-Instruct-v0.2-GPTQ as a backbone to predict labels. We compare the vanilla version of Mistral-7B-Instruct-v0.2-GPTQ (Mistral) to the ones that combine the enrichment generated by FLEX v0.2 (Mistral + FLEX v0.2) and FLEX v0.3 (Mistral + FLEX v0.3), respectively.
DistilBERT is a compressed version of the BERT model, developed to be smaller, faster, and more efficient while maintaining strong language understanding capabilities. Its key innovations include the use of knowledge distillation during the pre-training phase, a novel triple-loss function that integrates language modeling, distillation, and cosine-distance losses, as well as the implementation of large-batch training with gradient accumulation and dynamic masking. Similarly, we also examine the enrichment generated by different Mistral versions, e.g., FLEX v0.2 and FLEX v0.3.

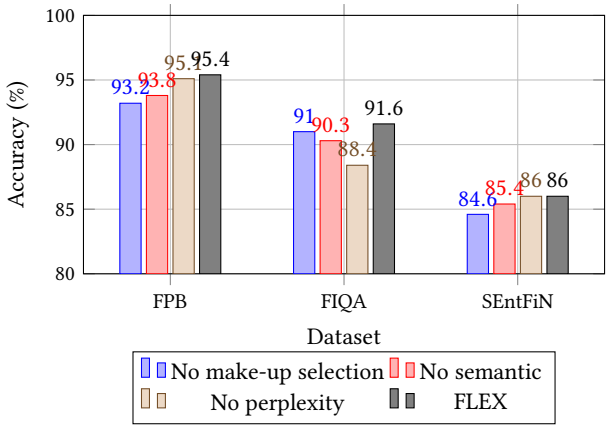
5 Results

The performance results of Mistral and DistilBERT models, both with and without the integration of FLEX versions, are summarized in Table 2. For decoder-only models (Mistral), the integration of FLEX v0.2 and v0.3 shows a clear improvement in both zero-shot and few-shot learning tasks across all datasets. In the zero-shot setting, Mistral combined with FLEX v0.2 achieves an average accuracy of 81.3%, compared to 75.7% without FLEX, demonstrating a substantial improvement. Mistral + FLEX v0.3 further increases accuracy to 83.6%, with notable gains particularly in the FPB and SEntFiN datasets. Similarly, in the few-shot setting, Mistral + FLEX v0.2 achieves an average accuracy of 84.4%, instead Mistral + FLEX v0.3 improves the accuracy to 86.0%.

Table 3: Case Study.

Setup	Content	Predicted	True
Original	In the second quarter of 2010 , Raute 's net loss narrowed to EUR 123,000 from EUR 1.5 million in the same period of 2009 .	NEGATIVE	POSITIVE
w/ FLEX	Raute's net loss was reduced from EUR 1.5 million in the second quarter of 2009 to EUR 123,000 in the same quarter of 2010, reflecting a notable improvement in the company's financial situation.	POSITIVE	
Original	$AAPL short some 592.49	POSITIVE	NEGATIVE
w/ FLEX	Shorting 592.49 shares of $AAPL stock is a bearish bet on the stock, with the expectation that its price will decrease in the near term and generate a profit from the subsequent decline.	NEGATIVE	
Original	3rd red day in a row ? $TSLA	POSITIVE	NEGATIVE
w/ FLEX	A third red day in a row for $TSLA's stock price indicates a bearish trend.	NEGATIVE	
Original	EXL beats profit estimates, cuts sales outlook	NEGATIVE	POSITIVE
w/ FLEX	EXL beats profit estimates, signifying that the company's earnings per share for the quarter were higher than the consensus forecast.	POSITIVE	
Original	Religare Finvest NCD issue oversubscribed 1.31 times	NEGATIVE	POSITIVE
w/ FLEX	Religare Finvest NCD issue oversubscribed 1.31 times, meaning the demand for the non-convertible debentures exceeded the supply by 31%.	POSITIVE	
Original	The solid fuel is heated before sludge is mixed therein .	POSITIVE	NEUTRAL
w/ FLEX	The solid fuel is heated before sludge is mixed therein, a standard practice to ensure the fuel is free of impurities before the sludge is added.	NEUTRAL	
Original	Homebuilders - $RYL breaking below support, watch this one. $SPY	POSITIVE	NEGATIVE
w/ FLEX	The homebuilding industry is experiencing a downturn, and the stock price of $RYL is reflecting this trend, as it has broken below a significant support level and is now trading below the $SPY index.	NEGATIVE	

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



For encoder-only models (DistilBERT), the results are particularly impressive in the fine-tuning setting, where the integration of FLEX v0.2 and v0.3 significantly boosts performance. DistilBERT alone achieves an average accuracy of 83.4%, but with FLEX v0.2, this increases to 87.2%, with the most substantial improvement observed in the FiQA dataset, where accuracy rises from 71% to 82.6%. FLEX v0.3 further enhances the model's accuracy to 91.0%, with gains across all datasets, including a notable jump in FiQA accuracy to 91.6%. These results indicate that the FLEX framework, particularly version 0.3, is highly effective in improving fine-tuning performance for financial sentiment analysis tasks, enabling models to better capture domain-specific nuances in complex financial data.

In summary, we can observe that FLEX can provide positive utilities across all the evaluation setups.

5.1 Ablation Study

We conducted an ablation study to enhance the robustness of our model and underscore the contribution of each component. Specifically, we tested each dataset using three distinct approaches, each designed to illustrate the importance of individual elements of our method. The experiments were performed 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 Makeup Selection:** Sentences are enriched directly using the LLM[4] without the makeup selection phase, meaning the LLM produces only one candidate.
- **No Semantic:** The LLM generates candidates as described in Sec.3, but the selection is based solely on the perplexity criterion, choosing the enriched sentence with the lowest perplexity value.
- **No Perplexity:** The LLM produces candidates as outlined in Sec.3, but the selection is made solely using the semantic criterion, choosing the enriched sentence with the highest cosine similarity to the original sentence.

Fig. 2 first highlights the value of generating multiple candidates rather than relying solely on the LLM's output. Additionally, the results reveal the significant impact of both components in our candidate selection criteria. The notable differences in trends between FPB, SEntFiN, and FIQA can be attributed to the varying sizes of these datasets. The FIQA dataset contains fewer documents compared to the other two, making longer sentences beneficial for the encoder during the fine-tuning training phase. This explains why the model without restrictions achieves the second-best performance.

Notably, the lowest performance is observed in the "No perplexity" solution, as it relies solely on semantic similarity with the original sentence, inherently resulting in shorter sentences. Nevertheless, the higher-quality sentences resulting from our selection criteria yield the highest accuracy. This can be explained by the limitations of using perplexity or semantic similarity alone. Perplexity by itself does not consider the similarity to the original sentence, which is crucial as it carries the sentiment label. Conversely, relying solely on semantic similarity risks tying the enriched sentence too closely to the original one, without adding meaningful content. For this reason, our approach uses perplexity to ensure that the new sentence is more predictable than the original while maintaining relevant semantic connections.

5.2 Case Study

In this section, we conduct case studies on the three datasets employed to qualitatively illustrate how our method is able to provide additional information that is useful for more accurate predictions.

As shown in Table 3, FLEX can paraphrase the original sentence such that the expression is more straightforward while also fitting to the financial domain. For instance, in the first example, our method reworded *"net loss narrowed"* to *"net loss was reduced"*; and in the second example, *"short"* to *"shorting shares of stock"*. The reworded expressions are easier to understand to both human and the machine, as indicated by the predictions with and without FLEX.

From the results, we can also observe that FLEX is able to correctly interpret and elaborate on the financial phenomena described in the given input, strengthening the intended sentiment of the original text. For instance, in the first example, FLEX further explained the stated fact by adding *"reflecting a notable improvement ..."*; in the second example, shorting shares of stock is interpreted a *"bearish bet"* and *"with the expectation that its price will decrease ..."*; in the third, *"a third red day in a row"* is elaborated with *"indicates a bearish trend"*; in the fourth, *"beats profit estimates"* is expanded as *"earnings ... were higher than the consensus forecast"*; and in the fifth, *"oversubscribe"* is elaborated to emphasize *"demand ... exceeded the supply"*. It can be seen that the elaborations are not only true to the original meanings, but also contain explicit sentiment indicators that aid the machine's prediction.

On the other hand, FLEX is also capable of expanding on facts unrelated to financial sentiment. For instance, in the sixth example, it added an explanation that the stated phenomenon in the original sentence is a standard practice in handling fuel, which helped the machine classify it as neutral in the financial context.

Furthermore, from the last example, it can be observed that FLEX is also capable of completely rewriting a sentence when the input is too informally worded. The rewritten sentence is true to the original meaning, i.e. *"RYL breaking below support"*, while being clearer and more comprehensible for both humans and machines.

6 Conclusion

In this work, we proposed an FSA approach that enhances the original input by incorporating semantic enrichment and perplexity measures, enabling the predictive model to make more accurate label predictions, based on a more comprehensive and coherent input.

Experimental results validate the effectiveness of our method across various setups, including zero-shot, few-shot, and fine-tuning frameworks. The ablation study further confirms the utility of the introduced semantic and perplexity checks. Future work can examine the enhancement of the method from the perspective of syntax [57] and pragmatics [31].

References

[1] Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063* (2019).

[2] 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)*. 675–718.

[3] 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*. 3417–3419.

[4] Erik Cambria. 2024. Understanding Natural Language Understanding. *Springer, ISBN 978-3-031-73973-6* (2024).

[5] Erik Cambria, Rui Mao, Melvin Chen, Zhaoxia Wang, and Seng-Beng Ho. 2023. Seven Pillars for the Future of Artificial Intelligence. *IEEE Intelligent Systems* 38, 6 (2023), 62–69.

[6] Erik Cambria, Xulang Zhang, Rui Mao, Melvin Chen, and Kenneth Kwok. 2024. SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing. In *Proceedings of International Conference on Human-Computer Interaction (HCII)*. Washington DC, USA.

[7] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[8] 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)*. IEEE, 1–10.

[9] 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)*. IEEE, 1127–1134.

[10] Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2023. FinSenticNet: A Concept-Level Lexicon for Financial Sentiment Analysis. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*. 109–114.

[11] Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. An Evaluation of Reasoning Capabilities of Large Language Models in Financial Sentiment Analysis. In *IEEE Conference on Artificial Intelligence (IEEE CAI)*. Singapore, 189–194.

[12] Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. Financial Sentiment Analysis: Techniques and Applications. *Comput. Surveys* 56, 9 (2024), 1–42. https://doi.org/10.1145/3649451

[13] 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* (2023).

[14] 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* (2023).

[15] 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*. 39–48.

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

[17] Jinhyuk Lee, Alexander Wettig, and Danqi Chen. 2021. Phrase Retrieval Learns Passage Retrieval, Too. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 3661–3672.

[18] 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 (2020), 9459–9474.

[19] 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*. 408–422.

[20] Qian Liu, Rui Mao, Xiubo Geng, and Erik Cambria. 2023. Semantic Matching in Machine Reading Comprehension: An Empirical Study. *Information Processing*

and Management 60, 2 (2023), 103145.
[21] Ruijun Liu, Yuqian Shi, Changjiang Ji, and Ming Jia. 2019. A survey of sentiment analysis based on transfer learning. *IEEE access* 7 (2019), 85401–85412.
[22] Shangqing Liu, Yu Chen, Xiaofei Xie, Jingkai Siow, and Yang Lie. 2021. Retrieval-augmented generation for code summarization via hybrid GNN.(2021). In *Proceedings of the Ninth International Conference on Learning Representations: ICLR*. 4–8.
[23] Yu Ma, Rui Mao, Qika Lin, Peng Wu, and Erik Cambria. 2023. Multi-source aggregated classification for stock price movement prediction. *Information Fusion* 91 (2023), 515–528.
[24] 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 (2024), 102165.
[25] 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*. 1941–1942.
[26] Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Filippo Pallucchini. 2024. SeNSe: embedding alignment via semantic anchors selection. *International Journal of Data Science and Analytics* (2024), 1–15.
[27] 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 (2014), 782–796.
[28] Rohan Manro, Rui Mao, Liza Dahiya, Yu Ma, and Erik Cambria. 2024. A Cognitive Analysis of CEO Speeches and Their Effects on Stock Markets. In *Proceedings of the 5th International Conference on Financial Technology (ICFT)*. Singapore.
[29] Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024. GPTEval: A Survey on Assessments of ChatGPT and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, Torino, Italia, 7844–7866.
[30] Rui Mao, Kelvin Du, Yu Ma, Luyao Zhu, and Erik Cambria. 2023. Discovering the cognition behind language: Financial metaphor analysis with MetaPro. In *2023 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1211–1216.
[31] Rui Mao, Mengshi Ge, Sooji Han, Wei Li, Kai He, Luyao Zhu, and Erik Cambria. 2025. A survey on pragmatic processing techniques. *Information Fusion* 114 (2025), 102712.
[32] Rui Mao, Kai He, Claudia Beth Ong, Qian Liu, and Erik Cambria. 2024. MetaPro 2.0: Computational Metaphor Processing on the Effectiveness of Anomalous Language Modeling. In *Findings of the Association for Computational Linguistics: ACL*. Association for Computational Linguistics, Bangkok, Thailand, 9891–9908.
[33] Rui Mao, Kai He, Xulang Zhang, Guanyi Chen, Jinjie Ni, Zonglin Yang, and Erik Cambria. 2024. A Survey on Semantic Processing Techniques. *Information Fusion* 101 (2024), 101988.
[34] Rui Mao, Xiao Li, Mengshi Ge, and Erik Cambria. 2022. MetaPro: A computational metaphor processing model for text pre-processing. *Information Fusion* 86-87 (2022), 30–43.
[35] Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The Biases of Pre-Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection. *IEEE Transactions on Affective Computing* 14, 3 (2023), 1743–1753. https://doi.org/10.1109/TAFFC.2022.3204972
[36] 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)*. 4089–4100.
[37] 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 (2020), 131662–131682.
[38] 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*. 12448–12460.
[39] 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* (2023).
[40] 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 (2022), 27730–27744.
[41] 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*. 2719–2734.
[42] Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. 2023. AART: AI-Assisted Red-Teaming with Diverse Data Generation for New LLM-powered Applications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 380–395.
[43] 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*.
[44] 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 (2022), 1314–1335.
[45] 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 (2018), 1–25.
[46] 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* (2022).
[47] 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 (2015), 4999–5010.
[48] 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)*. 563–570.
[49] 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* (2023).
[50] Frank Xing. 2024. Designing Heterogeneous LLM Agents for Financial Sentiment Analysis. *arXiv preprint arXiv:2401.05799* (2024).
[51] Frank Z Xing, Erik Cambria, and Roy E Welsch. 2018. Natural language based financial forecasting: a survey. *Artificial Intelligence Review* 50, 1 (2018), 49–73.
[52] Frank Z Xing, Filippo Pallucchini, and Erik Cambria. 2019. Cognitive-inspired domain adaptation of sentiment lexicons. *Information Processing & Management* 56, 3 (2019), 554–564.
[53] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. *FinLLM at IJCAI* (2023).
[54] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097* (2020).
[55] Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, S Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than Retrieve: Large Language Models are Strong Context Generators. In *International Conference on Learning Representations*.
[56] 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*. 349–356.
[57] Xulang Zhang, Rui Mao, and Erik Cambria. 2023. A Survey on Syntactic Processing Techniques. *Artificial Intelligence Review* 56 (2023), 5645–5728.
[58] Xulang Zhang, Rui Mao, and Erik Cambria. 2024. Multilingual Emotion Recognition: Discovering the Variations of Lexical Semantics between Languages. In *2024 International Joint Conference on Neural Networks (IJCNN)*. Yokohama, Japan.
[59] Xulang Zhang, Rui Mao, and Erik Cambria. 2024. SenticVec: Toward Robust and Human-Centric Neurosymbolic Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL*. Association for Computational Linguistics, Bangkok, Thailand, 4851–4863. https://aclanthology.org/2024.findings-acl.289
[60] Xulang Zhang, Rui Mao, Kai He, and Erik Cambria. 2023. Neuro-symbolic sentiment analysis with dynamic word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 8772–8783.
[61] Luyao Zhu, Rui Mao, Erik Cambria, and Bernard J. Jansen. 2024. Neurosymbolic AI for Personalized Sentiment Analysis. In *Proceedings of International Conference on Human-Computer Interaction (HCII)*. Washington DC, USA.