Sentiment Spin: Attacking Financial Sentiment with GPT-3

Markus Leippold
University of Zurich and Swiss Finance Institute

Sentiment Spin: Attacking Financial Sentiment with GPT-3

Markus Leippold[*]

February 2, 2023

Abstract

The use of dictionaries in financial sentiment analysis and other financial and economic applications remains widespread because keyword-based methods appear more transparent and explainable than more advanced techniques commonly used in computer science. However, this paper demonstrates the vulnerability of using dictionaries by exploiting the eloquence of GPT-3, a sophisticated transformer model, to generate successful adversarial attacks on keyword-based approaches with a success rate close to 99% for negative sentences in the financial phrase base, a well-known human-annotated database for financial sentiment analysis. In contrast, more advanced methods, such as those using context-aware approaches like BERT, remain robust.

Version: February 2, 2023
JEL-Code: G2, G38, C8, M48
Keywords: Sentiment analysis in financial markets, keyword-based approach, FinBERT, GPT-3.

[*] Department of Banking and Finance, University of Zurich, Switzerland, and the Swiss Finance Institute (SFI), Plattenstrasse 14, 8032 Switzerland. Email: markus.leippold@bf.uzh.ch.

1 Introduction

Recently, Cao et al. (2022) suggest that increasing AI readership encourages firms to craft filings tailored to machine parsing and processing. Their research is the first to recognize the feedback effect of how firms adjust their language because they know machines are listening.[1] These effects could potentially lead to manipulation and collusion. In particular, the authors find that companies actively started managing the sentiment and tone of their disclosures to produce desirable outcomes from algorithms. This is all the more of a concern as sentiment can strongly impact returns and volatility of financial assets, moving prices away from their fundamentals.[2] In light of recent developments in large language models, this paper investigates whether financial texts can be algorithmically manipulated so that AI readers (machines) can be deceived by adversarial attacks in their financial sentiment analysis.

The finance literature is still dominated by keyword-based approaches to sentiment analysis of financial texts, which are predominantly based on the Loughran and McDonald (2011) dictionary.[3] In these rule-based approaches, the presence or absence of certain words or phrases in a text is detected to determine the overall sentiment by counting how often these words occur in the text. Alternatively, this can be done automatically by using natural language processing (NLP) techniques to identify and extract the relevant keywords from the text. One advantage of keyword-based approaches is that they are transparent, relatively easy to implement, and can be applied to a variety of financial texts, including annual reports, press releases, and social media posts. Another argument is that they are

[1] The first version of their paper, published in 2020, refers to AI systems as predominantly based on keyword-based methods. Their version from December 2022 also includes a BERT Devlin et al. (2018) model in their analysis.

[2] There is a vast literature on this subject in finance. See, e.g., Baker and Wurgler (2007), among many others. Already Engle and Ng (1993) were analyzing the asymmetric impact of (negative and positive) news on volatility. A recent overview on defining and measuring market sentiment is given, e.g., in Aggarwal (2022).

[3] As of January 2023, the paper of Loughran and McDonald (2011) has been cited more than 4,276 times by researchers. Moreover, their word list has been adopted for the WRDS SEC Sentiment Data.

easier to interpret and require fewer computational resources than other, more modern NLP approaches that rely almost exclusively on deep neural networks.

However, keyword-based approaches may also have some limitations. For example, they cannot accurately capture the context or tone of the text and are not flexible enough to detect sarcasm or irony. As a result, they do not always provide reliable or accurate results, especially when applied to more complex or nuanced financial texts. Indeed, recent literature, also from the non-financial domain,[4] indicates that the performance of dictionaries is not satisfactory, which can be mostly attributed to the missing context of words. The same verdict applies to financial dictionaries (e.g. Frankel et al. 2022). Despite these findings, dictionaries remain the dominant approach in academic finance, with new dictionaries emerging that promise to improve upon previous methods.[5]

The goal of this paper is not to revisit the debate about whether keyword-based or more advanced methods are better for sentiment classification in finance, as this has been adequately addressed in other studies.[6] Instead, this paper focuses on investigating the vulnerability of sentiment classification approaches to adversarial attacks using large language models such as GPT-3 (Generative Pre-trained Transformer 3). Adversarial attacks on NLP systems involve introducing subtle modifications to the original text that are imperceptible to humans yet cause the model to generate incorrect predictions.[7] These attacks have the potential to undermine the reliability and accuracy of sentiment classification.

[4] See, e.g., Boukes et al. (2020), Hartmann et al. (2022), Van Atteveldt et al. (2021) They observe, for instance, that off-the-shelf sentiment dictionaries rarely work well beyond the text genres they were developed from, with agreement generally close to a chance agreement, and correlations between dictionaries were also low. For comparison in the context of climate change, also see Bingler et al. (2022), Webersinke et al. (2021).

[5] Examples include Benchimol et al. (2021), De Franco et al. (2022), Huang et al. (2022), Jiang et al. (2019), Katsafados et al. (2021), Ma and Xu (2021), among others. In addition, new dictionaries emerge (Garcia et al. 2023), with the promise to outperform previous dictionaries.

[6] See, e.g., Frankel et al. (2022). They show how machine-learning methods outperform keyword-based approaches in the financial context. Moreover, a survey is given in Zhu et al. (2022). Lastly, a highly comprehensive overview of different transformed models is given in Mishev et al. (2020).

[7] Research on adversarial attacks on neural networks began in computer vision (Szegedy et al. 2013, Goodfellow et al. 2014). In contrast to computer vision, adversarial examples in NLP cannot be imperceptible, as discrete characters or words must be replaced.

Previous NLP research on adversarial attacks has primarily used rule-based methods to create negative examples by replacing words with synonyms. Unfortunately, these methods can lead to unnatural and confusing changes in the text that people can easily detect. However, recent advances in NLP have been tremendous, with large language models (LLMs) being developed that show impressive results in language generation. This has led to renewed interest in adversarial attacks in the NLP field, with a variety of different methods proposed, such as, e.g., Papernot et al. (2016), Alzantot et al. (2018), Jin et al. (2020), Garg and Ramakrishnan (2020), Wang et al. (2021).[8] Yet, recent studies have shown that even these highly advanced methods often do not preserve the semantics of the original text, contrary to what has often been claimed in previous work. In particular, Morris et al. (2020) and Hauser et al. (2021) have shown that most perturbations made by adversarial attacks do not preserve the semantics of the original text.[9] Therefore, to construct adversarial attacks, I will rely on the eloquence of GPT-3 (Brown et al. 2020), the current state-of-the-art tool for generating contextually and semantically correct text.

2 NLP Methods

I will now briefly give an overview of the methods used in this study: the keyword-based methodology, which is based on a financial lexicon, and a specific transformer-based approach, which provides state-of-the-art sentiment classification for financial texts. Both methodologies are freely available tools and open-source. For text generation, I use GPT-3, which can only be accessed via an API at the time of this writing.

[8] For example, in Garg and Ramakrishnan (2020), the authors propose a new approach to generating adversarial examples using BERT-based tokens that is able to retain the semantic meaning of the sentence in natural-looking sentences. Different from other pipelines, it utilizes dynamic embedding vector spaces rather than fixed embedding vector space.

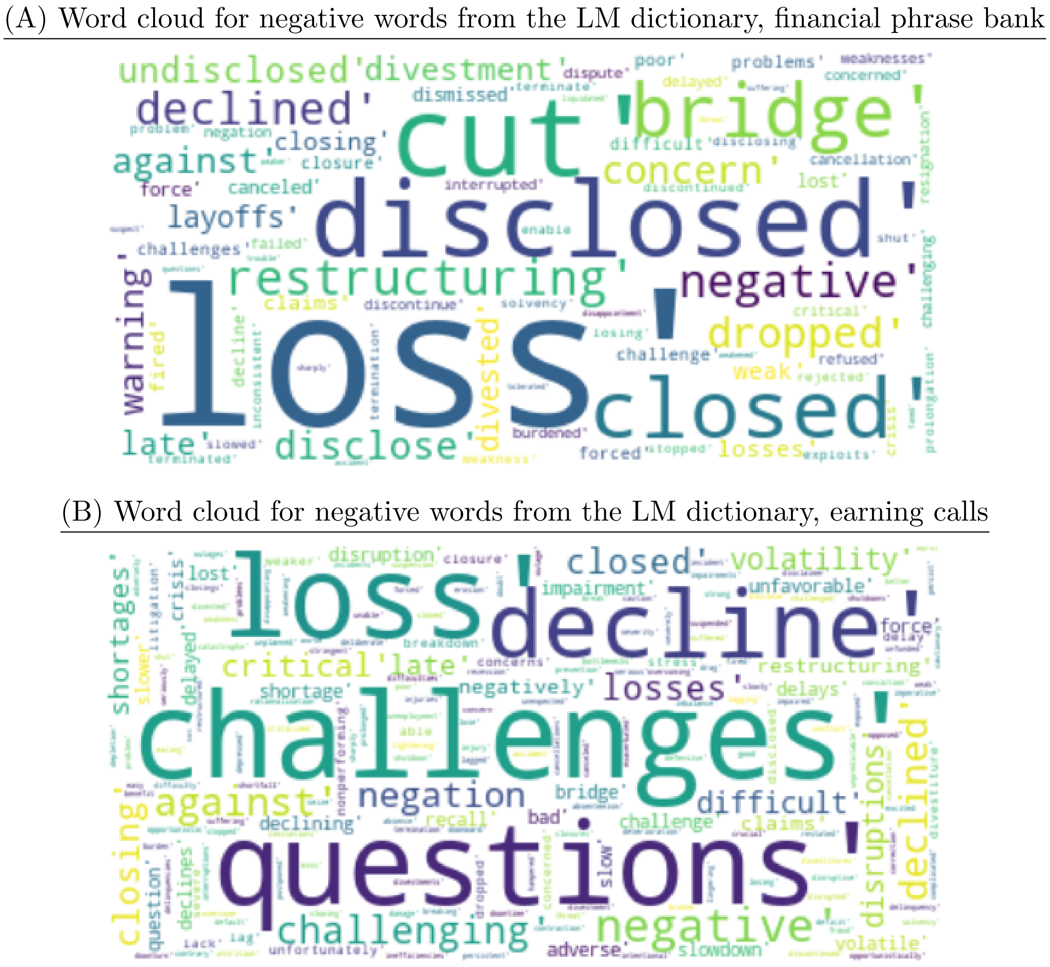
[9] The authors find that between 96% and 99% of the analyzed attacks do not preserve semantics. They conclude that BERT is much more robust than research on attacks suggests.

2.1 Financial dictionaries

The Henry (2008) Financial Dictionary (HFD) and the Loughran and McDonald (2011) dictionary (LM) are the most prominent financial dictionaries. HFD is the first dictionary specifically designed for the financial domain and contains 104 positive and 85 negative words. It is used to assess the tone of earnings press releases, which are a key element of communication between firms and investors. HFD has been frequently used for financial sentiment analysis, but its limited number of words and low coverage are weaknesses. In contrast, the LM dictionary is a sentiment word list compiled from annual reports released by firms and includes 354 positive, 2,355 negative, 297 uncertainty, 904 litigious, 19 strong modal, 27 weak modal, and 184 constraining words.

LM is the most widely used finance domain lexicon that we are aware of. The paper's provision of a specialized finance dictionary, containing both positive and negative connotations and words that offer information regarding prospects and uncertainty, has influenced sentiment analysis algorithms in the industry and academia. Given that the LM dictionary has become the predominant dictionary used in lexicographic algorithms for sentiment calibration, we focus on the LM in the subsequent analysis. In addition, the HFD is much more vulnerable to attacks due to the limited number of words.

Figure 1 shows the word cloud of words that appear in sentences for which the LM dictionary would attach a negative sentiment. The sentences are taken from the financial phrase bank (Malo et al. 2013) (4,846 human-annotated sentences from news headlines) and from a subsample of earning calls from 2022 (500,000 sentences). A dictionary-based approach would take negative words and positive words in each sentence and aggregate them to obtain a sentence-level score for the sentiment eventually. It is clear from Figure 1 that the diversity of negative words from LM in the bank phrase bank is much lower than in the larger dataset generated from the earning calls.



(A) Word cloud for negative words from the LM dictionary, financial phrase bank

(B) Word cloud for negative words from the LM dictionary, earning calls

Figure 1: In Panel A, we plot the word cloud for the negative words from the LM dictionary as they appear in the financial phrase bank (4,846 sentences). In Panel B, we plot the corresponding word cloud for sentences taken out of a subsample of earning calls from 2022 (500,000 sentences).

2.2 FinBERT

Recent advancements in architecture and training methods have significantly improved the way deep learning models are able to learn general language to perform a variety of different downstream tasks. With self-attention and token masking, the transformer architecture of BERT (Devlin et al. 2018) and other deep neural networks allows a truly bi-directional contextual relationship to be learned. The transformer architecture is a key component of these models, a type of neural network architecture particularly well-suited for NLP tasks, such as language understanding and generation. The underlying attention mechanism allows the model to focus selectively on different parts of the input when processing it.[10]

Often, these large language models are trained on a large corpus of general text and can further be improved if pretrained on domain-specific language. For this reason, various versions of BERT models were introduced specifically for the financial domain and christened FinBERT (Araci 2019, Liu et al. 2020, Yang et al. 2020, Hazourli 2022). For our analysis, we use the version of FinBERT as introduced by Araci (2019).[11] The transformer architecture allows FinBERT to process input text more sophisticatedly by selectively focusing on different parts of the input. The fine-tuning of FinBERT on financial corpus allows the model to understand better the financial domain, which can be useful for various classification tasks.

[10] The transformer architecture is composed of an encoder and a decoder. The encoder takes in a sequence of tokens (e.g., words or subwords) and generates a set of hidden representations, one for each token in the sequence. These hidden representations capture the meaning of the input text in a high-dimensional space. The decoder then takes in the hidden representations and generates a set of output tokens, which can be used for a variety of NLP tasks, such as language understanding, language generation, and text classification.

[11] In some preliminary analysis, I find that the FinBERT-version of (Araci 2019) outperforms all others. Therefore, I only consider that model for my experiments. Araci (2019) use the original BERT model and further pre-trains it on a subset of the TRC2 corpus, a collection of 1.8M news articles published by Reuters between 2008 and 2010, by filtering for keywords related to finance. Moreover, their model is already fine-tuned on sentiment analysis, and I use their sentiment classifier downloaded from Hugging Face (`https://huggingface.co/ProsusAI/finbert`).

2.3 GPT-3

GPT-3 is a language generation model developed by OpenAI based on a transformer architecture (Brown et al. 2020) with a decoder only but no encoder. Decoders are created for text generation, which makes them particularly suitable for tasks like machine translation, summarisation, and abstractive question-answering, but less so for classification tasks like sentiment analysis. GPT-3 is trained on a massive dataset of internet text and can generate human-like text, complete tasks such as translation and summarization, and even write code. With 175 billion parameters, it is considered one of the most advanced language generation models currently available and has been used in a variety of applications, including language translation and text generation. Given its strength as a pure neural decoder network, I will use GPT-3 for the text generation required for adversarial attacks. However, I will also present some interesting results on sentiment classification (which is not considered a particularly difficult task in NLP).

3 Data

Luo et al. (2018) categorize financial sentiment indicators into market-derived and human-annotated sentiment. The market-derived sentiment was calculated from market dynamics, such as price movement and trading volume. A drawback of this approach is that the results may contain noise from other sources. For this reason, I rely on subjective human annotated sentiments annotated by professionals (Malo et al. 2013) or investors themselves (**?**). For my experiments, I use the former database, the well-known financial phrase bank developed in (Malo et al. 2013), which contains 4,837 English news headlines of companies listed on the OMX Helsinki exchange. It was annotated by a team of 16 people with a background in finance, economics, or accounting, who classified each sentence based on their emotional tones as either positive, negative, or neutral. The total number of sentences corresponds to the instances of an inner-annotator agreement larger than 50%. From these sentences, there are 2,264 sentences with a 100% agreement. For my experiment, we will use the full database with 4,846 sentences, consisting of 604 negative, 2,879 neutral, and 1,363 positive sentences. Examples of the sentences and their labels from human annotations are given in Table B.1. I exclusively focus on sentences that human annotators have negatively annotated since our goal in adversarial attacks is to turn the negative sentiment into neutral or positive. For the second part of my experiments, I will use unlabelled data from earning calls.

4 Sentiment analysis on the original data

Before we move on to creating adversarial attacks, it is instructive to see how well the various sentiment analysis methods perform on the original data. Since Mishev et al. (2020), it is clear that FinBERT will outperform the LM dictionary approach. However, given the recent hype around GPT-3 (and chatGPT), I will also include GPT-3 with zero-shot and few-shot learning.[12]

Table 1 reports the different performance metrics for all the models. It turns out that FinBERT outperforms not only the keyword-based approach but also the GPT-3 few-shot learners by a considerable margin. The weighted average F1-score is 0.89 for FinBERT, while for the best GPT-3 model, the F1-score is 0.74. The keyword-based approach lags far behind with an F1 score of 0.60. Hence, in what follows, I will not rely on GPT-3 to perform the sentiment analysis. Instead, I will only consider FinBERT and the more traditional keyword-based approach. However, GPT-3 will help me in formulating adversarial attacks on these two methods.

Figure 2 gives the confusion matrix for the keyword-based approach and the FinBERT with the true labels on the y-axis and the predicted labels on the x-axis. Clearly, the

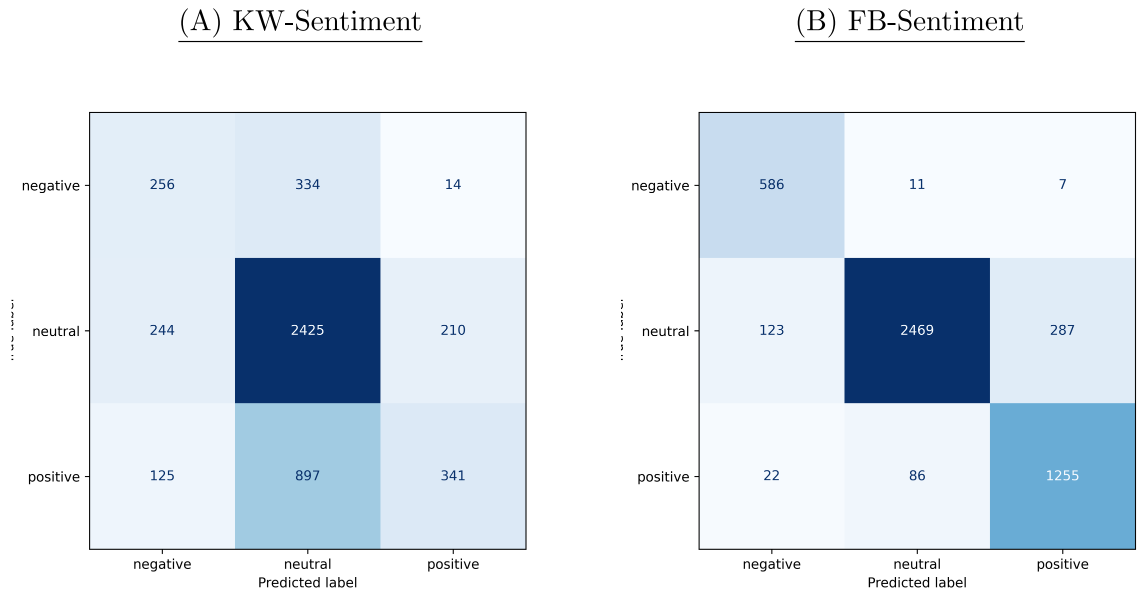
[12] However, it can be expected that GPT-3 might struggle with the classification task as it is a pure decoder model. For this paper, I use the version Davinci-003. In unreported results, I also used Davinci-001 and the classification results were disappointingly close to random, with F1-scores around 0.5. Obviously, the model has been improved considerably.

	Keyword			FinBERT		
	precision	recall	f1-score	precision	recall	f1-score
negative	0.410	0.424	0.417	0.802	0.970	**0.878**
neutral	0.663	0.842	0.742	0.962	0.858	**0.907**
positive	0.604	0.250	0.354	0.810	0.920	**0.862**
macro avg	0.559	0.505	0.504	0.858	0.916	**0.882**
weighted avg	0.615	0.624	0.592	0.899	0.889	**0.890**
	GPT-3 (3-shots)			GPT-3 (6-shots)		
	precision	recall	f1-score	precision	recall	f1-score
negative	0.890	0.457	0.604	0.894	0.710	0.792
neutral	0.777	0.794	0.786	0.865	0.617	0.720
positive	0.647	0.756	0.698	0.547	0.913	0.684
macro avg	0.772	0.669	0.696	0.769	0.747	0.732
weighted avg	0.755	0.742	0.738	0.780	0.712	0.719

Table 1: The table reports different performance measures for the different algorithms on the sentiment task.

keyword-based approach struggles to attach positive sentiment to truly positive sentences. This might be caused by the fact that the LM dictionary has much fewer positive (354) than negative words (2,355). However, the keyword-based approach also struggles to identify truly negative sentences (only 256 out of 604), while FinBERT is able to do a reasonably good job (586 out of 604).

One could argue that the keyword-based approach will perform much better when applied to sentences where human annotators fully agree on the sentiment. However, it turns out that this is not the case. In Table 2, we report the corresponding results when we restrict the analysis to this subset. While the weighted average F1-score jumps to 0.97, the keyword-based approach only generates a small and negligible increase from 0.592 to 0.599. Despite this result, the keyword-based approach is still dominant in the academic literature on financial sentiment analysis. To show that this approach should be abandoned, especially given recent advances in natural language processing, I will show how the keyword-based approach can suffer severely from adversarial attacks. There are many different ways to design such adversarial attacks. However, with the arrival of generative


(A) KW-Sentiment

(B) FB-Sentiment

Figure 2: Confusion matrix for the financial sentiment analysis on the financial phrase bank. Panel A shows the results for the keyword-based approach; Panel B shows the result for FinBERT. Table 1 gives a detailed summary of the performance metrics. The annotated true labels are on the y-axis, and the predicted labels are on the x-axis

models like GPT-3, a very powerful pipeline can be defined with minimal effort.

5 Manipulating sentiment with GPT-3

GPT-3 is, among other tasks, a great tool for sentence completion and text generation, and hence perfectly suited to manipulate sentences from the financial phrase bank. By using GPT-3, I try to overcome some of the problems mentioned in the previous literature on adversarial attacks, namely that the manipulated sentences give semantically non-meaningful results and can easily be spotted by humans (Hauser et al. 2021). Moreover, GPT-3 dramatically simplifies the pipeline for adversarial attacks.[13]

[13] For instance, similar to previous literature (Jin et al. 2020), I tried using BERT embeddings to generate synonyms by averaging over all the attention layers. However, the results were not good, so I switched back to the use of GPT-3.

	Keyword			FinBERT		
	precision	recall	f1-score	precision	recall	f1-score
negative	0.418	0.403	0.410	0.909	0.983	**0.945**
neutral	0.686	0.901	0.779	0.998	0.968	**0.982**
positive	0.637	0.163	0.260	0.947	0.975	**0.961**
accuracy	0.648	0.648	0.648	0.972	0.972	**0.972**
macro avg	0.580	0.489	0.483	0.951	0.976	**0.963**
weighted avg	0.638	0.648	0.599	0.973	0.972	**0.972**

Table 2: Performance of keyword-based approach and FinBERT on the subset of the financial phrase bank where annotators agree with 100%.

5.1 Strategy 1: Prompting GPT-3

In Strategy 1, I will follow a very direct approach in prompting GPT-3.[14] In particular, I will apply the code in Listing 1, given in Appendix A, to change the negative sentiment to neutral or positive. The intuition behind the code is quite simple. First, I ask GPT-3 to generate a list of potential synonyms that fit the context of the word(s) identified from a given sentence that is(are) also part of the LM dictionary. Then, I filter out all the suggested synonyms that appear in the negative words list in the dictionary. Lastly, I ask GPT-3 to rephrase the sentence such that it replaces the negative words of the LM dictionary with the suggested synonyms while respecting the context, the meaning, and the grammar of the sentence.

5.2 Strategy 2: Prompting GPT-3 given a predefined list of synonyms

For Strategy 2, I will first generate a new dictionary based on synonyms generated by GPT-3, which are not yet part of the LM dictionary. In particular, I ask GPT-3 to generate up to 100 synonyms for each negative word in the LM dictionary. At some point, GPT-3 becomes repetitive. Therefore, I only keep the unique words. In addition, I filter out all those words that are already part of the LM dictionary. In the second step, given a sentence with a negative word from the LM dictionary, I ask GPT-3 to rephrase the sentence by replacing

[14] I also tried few-shot learning. However, the results were not improving over zero-shot learning.

the negative word with a corresponding synonym from my newly generated dictionary. The disadvantage compared to Strategy 1 is that GPT-3 no longer has the context of the word for which it must generate the synonyms. By doing so, I risk that the rephrased sentence might not be semantically correct. However, given that I want to apply this method to a couple of thousands of sentences, I can considerably reduce the cost of using the GPT-3 API. The code snippets are provided in Listing 2 and 3 of Appendix A, together with the list of generated synonyms for all the negative words in the LM dictionary.

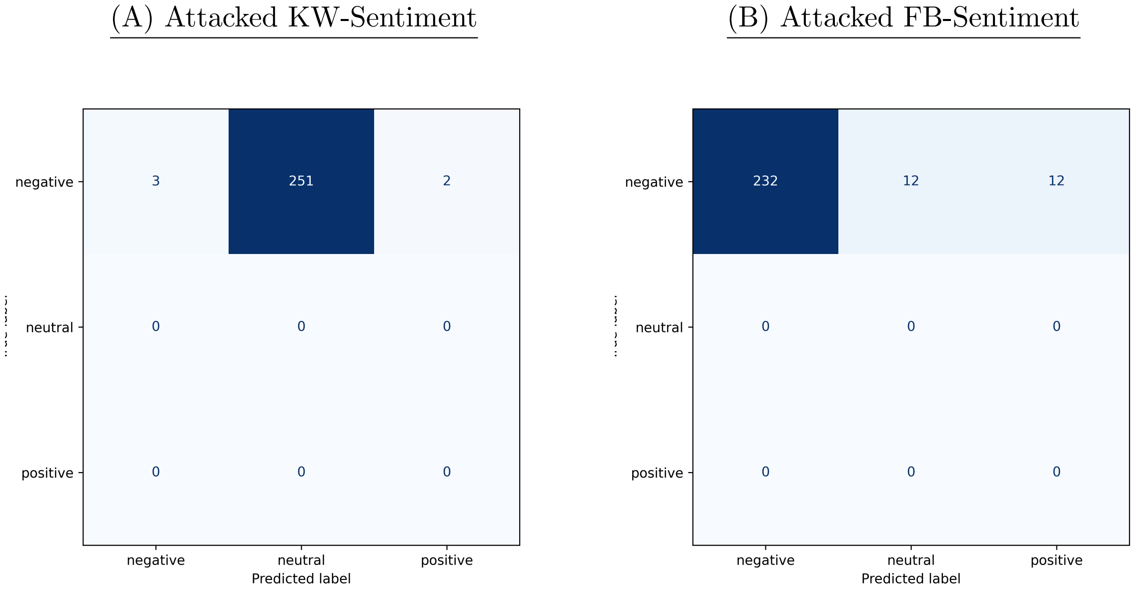
6 Experiments: Positive sentiment spin

What is the impact of the adversarial attacks on negative sentiments, i.e., can we find ways to spin the negative sentiment of a given sentence into a neutral sentiment (or even a positive one)? First, I attack the financial phrase dataset since this dataset contains annotated data using the strategies outlined in the section above. The keyword-based approach and FinBERT will then classify the attacked sentences. Then, I will also perform the same analysis on sentences taken from earning calls.

6.1 Manipulating sentiment in the financial phrase bank

In the financial phrase database, we can find 256 sentences annotators have labeled as 'negative,' and the keyword-based approach correctly assigns a negative sentiment. These sentences form the basis of my experiments. In the first experiment, I use Strategy 1, described in the previous section. Given the new sentences generated by GPT-3, I can assess the sentiment using the keyword-based approach and FinBERT.

Figure 3 display the results. From all the 256 sentences, only three survive the attack under the keyword-based approach, i.e., all other sentences obtain a neutral or positive sentiment. In contrast, FinBERT, which correctly assesses a negative sentiment for the original sentences in 98% of the cases, decreases its accuracy to only 91%. Hence, the



(A) Attacked KW-Sentiment

(B) Attacked FB-Sentiment

	Keyword			FinBERT		
	precision	recall/acc	f1-score	precision	recall/acc	f1-score
before attack	1.000	1.000	1.000	1.000	0.980	0.990
after attack	1.000	0.012	0.023	1.000	0.906	**0.951**

Figure 3: The figure shows the confusion matrix for the attacked sentences with negative sentiment. The sample consists only of sentences that originally have been correctly identified by the keyword-based approach with a negative sentiment. The table below reports the different performance measures. In this case, the recall is equal to the accuracy (acc) since we only have true positives. The accuracy of the attacked KW approach is reduced (from 100%) to 1.2%, while the accuracy of the attacked FB approach remains at 90.6%. On this subset, the accuracy of FB on the original sentences is 98%.

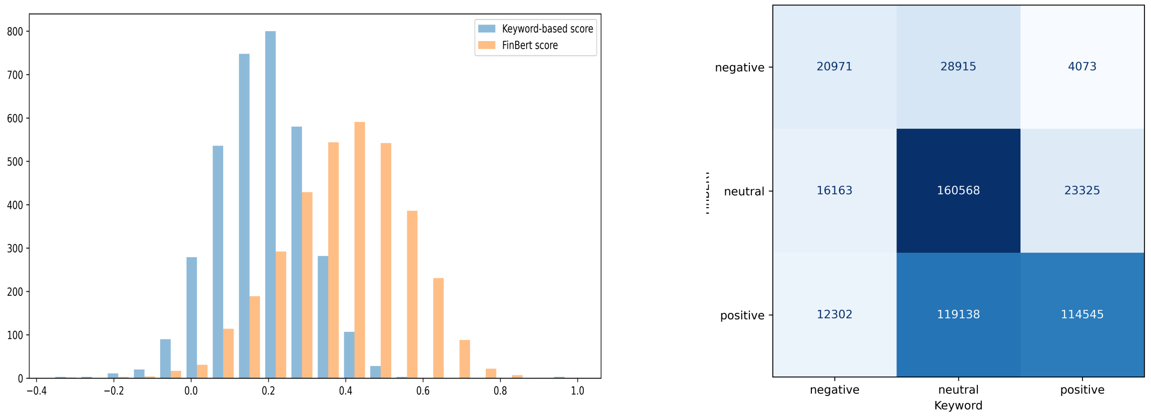
attack's success rate is at 99% for the keyword-based approach, while it is only at 7% for FinBERT. Table B.2 in Appendix B gives some examples of attacked sentences. In principle, the semantic quality of the sentences is high. However, GPT-3 struggles with generating synonyms for more frequent words like 'loss' (see Figure 1), which eventually results in a sentence that would look suspicious to a human (if that human knows that the sentence could have been potentially subject to an adversarial attack). Nevertheless, the overall quality of the adversarial attacks is quite high, given the simplicity of how these attacks are generated.

6.2 Manipulating sentiment in earning calls

For my second experiment, I use sentences from the presentation sections of earning calls in 2022. In Figure 4, Panel A, I have illustrated the distribution of sentiment scores calculated using a keyword-based method and FinBERT on 500,000 sentences from earning call transcripts in 2022. The scores represent the average sentiment across all sentences in the presentation section for a specific company. In Panel B, the confusion matrix for the corresponding sentiments is depicted. Since we do not have human-annotated sentences, we do not have a reference point or ground truth to compare against, unlike the financial phrase bank. The figure suggests that there is not much agreement between the sentiment scores of the keyword-based approach and FinBERT. Moreover, there seems to be a negative bias in how the keyword-based approach scores the sentiment of earning call sentences.

Figure 5 shows the distributions of sentiment scores computed using the keyword-based approach for 35,000 sentences from call transcripts in 2022.[15] The sample size of sentences with negative sentiment obtained by the keyword-based approach is 3,902. These sentences with negative sentiment were then attacked using GPT-3, specifically Strategy 2, as described above. The resulting sentiment score is then calculated as the average score for all

[15] I have reduced the number of sentences attacked by GPT-3 to 3,902 due to computational constraints. However, the results hold without loss of generality



(A) Original distribution of KW- and FB-Sentiment

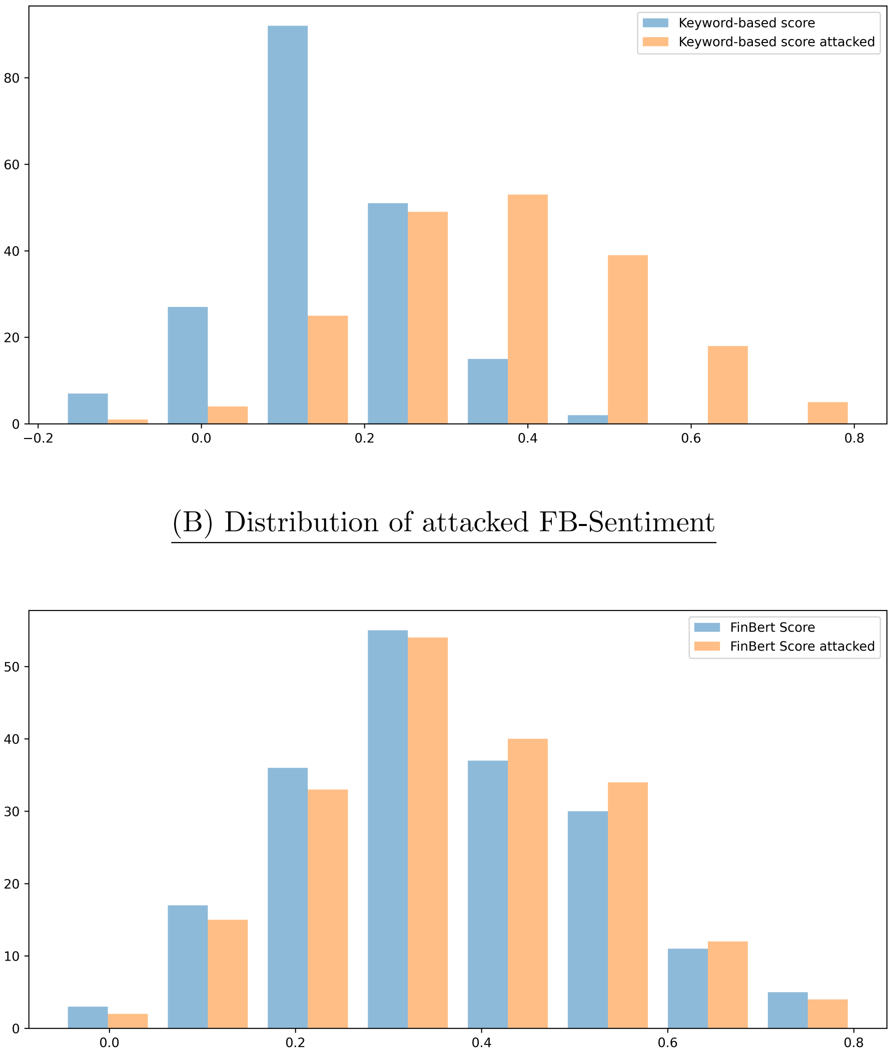
(B) Confusion Matrix

Figure 4: In Panel A, I plot the original distribution of the sentiment score on 500,000 sentences, calculated using a keyword-based approach and using FinBERT. taken from earning calls in 2022. The score is calculated as the mean sentiment over all the sentences in the presentation section for a given company. In Panel B, I plot the confusion matrix for the corresponding sentiments. Unlike the financial phrase bank, we do not have human-annotated sentences; therefore, we do not have a ground truth as a benchmark.

sentences in the presentation section for a given company. Panel B of Figure 5 shows the distribution of sentiment scores when computed with the FinBERT model, again for the original sentences and the sentences attacked with GPT-3. It is clear that attacking the keyword-based approach leads to a large difference in the distribution of sentiment scores between the companies. At the same time, the distribution of the attacked FinBERT classifier remains close to the original distribution. Some example sentences are given in Table B.3 in Appendix B, together with the verdicts of the keyword-based approach and FinBERT and my own (human) assessment.

7 Conclusion

Overall, my analysis shows the main problem of keyword-based approaches. They lack robustness, even if highly popular and broad dictionaries like the one of Loughran and



(A) Distribution of attacked KW-sentiment

(B) Distribution of attacked FB-Sentiment

Figure 5: In Panel A, I plot the distribution of the sentiment score on 35,000 sentences, calculated using a keyword-based approach, when the sentences with the negative sentiment of the keyword-based approach are attacked by GPT-3. The sample size of sentences with negative sentiment is 3,902. The earning calls are from 2022. The score is calculated as the mean sentiment over all the sentences in the presentation section for a given company. In Panel B, I plot the corresponding distributions when calculating the score using FinBERT.

McDonald (2011) are used. Hence, the same problem may persist even more so for domain-specific dictionaries. More involved methods, such as BERT, are more robust to attacks by a generative model like GPT-3. Nevertheless, there might be other ways for adversarial attacks. The approach taken in this paper is, on purpose, rather simple, showcasing the power of current NLP methods to generate adversarial attacks. More elaborated models could potentially make these adversarial models even more powerful. Therefore, this topic should be further explored, especially in light of the information overload that makes AI-based information processing increasingly indispensable.

References

Aggarwal, D.: 2022, Defining and measuring market sentiments: A review of the literature, *Qualitative Research in Financial Markets* **14**(2), 270–288.

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.-J., Srivastava, M. and Chang, K.-W.: 2018, Generating natural language adversarial examples, *arxiv preprint arxiv:1804.07998*.

Araci, D.: 2019, FinBERT: Financial sentiment analysis with pre-trained language models, *arxiv preprint arxiv:1908.10063*.

Baker, M. and Wurgler, J.: 2007, Investor sentiment in the stock market, *Journal of Economic Perspectives* **21**(2), 129–151.

Benchimol, J., Kazinnik, S. and Saadon, Y.: 2021, Federal reserve communication and the covid-19 pandemic, *Covid Economics* **79**, 218–256.

Bingler, J. A., Kraus, M., Leippold, M. and Webersinke, N.: 2022, Cheap talk in corporate climate commitments: The effectiveness of climate initiatives, *Available at SSRN 3998435* .

Boukes, M., Van de Velde, B., Araujo, T. and Vliegenthart, R.: 2020, What's the tone? Easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools, *Communication Methods and Measures* **14**(2), 83–104.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al.: 2020, Language models are few-shot learners, *Advances in neural information processing systems* **33**, 1877–1901.

Cao, S., Jiang, W., Yang, B. and Zhang, A. L.: 2022, How to talk when a machine is listening: Corporate disclosure in the age of ai, *SSRN Working Paper*.

De Franco, G., Shohfi, T., Xu, D. and Zhu, Z. V.: 2022, Fixed income conference calls, *Journal of Accounting and Economics* p. 101518.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: 2018, Bert: Pretraining of deep bidirectional transformers for language understanding, *arxiv preprint arxiv:1810.04805*.

Engle, R. F. and Ng, V. K.: 1993, Measuring and testing the impact of news on volatility, *The Journal of Finance* **48**(5), 1749–1778.

Frankel, R., Jennings, J. and Lee, J.: 2022, Disclosure sentiment: Machine learning vs. dictionary methods, *Management Science* **68**(7), 5514–5532.

Garcia, D., Hu, X. and Rohrer, M.: 2023, The colour of finance words, *Journal of Financial Economics* **147**(3), 525–549.

Garg, S. and Ramakrishnan, G.: 2020, Bae: BERT-based adversarial examples for text classification.

Goodfellow, I. J., Shlens, J. and Szegedy, C.: 2014, Explaining and harnessing adversarial examples, *arxiv preprint arxiv:1412.6572*.

Hartmann, J., Heitmann, M., Siebert, C. and Schamp, C.: 2022, More than a feeling: Accuracy and application of sentiment analysis, *International Journal of Research in Marketing* .

Hauser, J., Meng, Z., Pascual, D. and Wattenhofer, R.: 2021, Bert is robust! a case against synonym-based adversarial examples in text classification, *arxiv preprint arxiv:2109.07403*.

Hazourli, A.: 2022, Financialbert - a pretrained language model for financial text mining, *Technical report*.

Henry, E.: 2008, Are investors influenced by how earnings press releases are written?, *The Journal of Business Communication (1973)* **45**(4), 363–407.

Huang, J., Roberts, H. and Tan, E. K.: 2022, The media and ceo dominance, *International Review of Finance* **22**(1), 5–35.

Jiang, F., Lee, J., Martin, X. and Zhou, G.: 2019, Manager sentiment and stock returns, *Journal of Financial Economics* **132**(1), 126–149.

Jin, D., Jin, Z., Zhou, J. T. and Szolovits, P.: 2020, Is Bert really robust? a strong baseline for natural language attack on text classification and entailment, *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, pp. 8018–8025.

Katsafados, A. G., Androutsopoulos, I., Chalkidis, I., Fergadiotis, E., Leledakis, G. N. and Pyrgiotakis, E. G.: 2021, Using textual analysis to identify merger participants: Evidence from the us banking industry, *Finance Research Letters* **42**, 101949.

Liu, Z., Huang, D., Huang, K., Li, Z. and Zhao, J.: 2020, FinBERT: A pre-trained financial language representation model for financial text mining, *in* C. Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, International Joint Conferences on Artificial Intelligence Organization, pp. 4513–4519. Special Track on AI in FinTech.
URL: *https://doi.org/10.24963/ijcai.2020/622*

Loughran, T. and McDonald, B.: 2011, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, *The Journal of Finance* **66**(1), 35–65.

Luo, L., Ao, X., Pan, F., Wang, J., Zhao, T., Yu, N. and He, Q.: 2018, Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention., *IJCAI*, pp. 4244–4250.

Ma, Y. and Xu, L.: 2021, Major government customers and stock price crash risk, *Journal of Accounting and Public Policy* **40**(6), 106900.

Malo, P., Sinha, A., Takala, P., Korhonen, P. J. and Wallenius, J.: 2013, Good debt or bad debt: Detecting semantic orientations in economic texts, *CoRR* **abs/1307.5336**.
URL: *http://arxiv.org/abs/1307.5336*

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T. and Trajanov, D.: 2020, Evaluation of sentiment analysis in finance: from lexicons to transformers, *IEEE access* **8**, 131662–131682.

Morris, J. X., Lifland, E., Lanchantin, J., Ji, Y. and Qi, Y.: 2020, Reevaluating adversarial examples in natural language.

Papernot, N., McDaniel, P., Swami, A. and Harang, R.: 2016, Crafting adversarial input sequences for recurrent neural networks, *MILCOM 2016-2016 IEEE Military Communications Conference*, IEEE, pp. 49–54.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R.: 2013, Intriguing properties of neural networks, *arxiv preprint arxiv:1312.6199*.

Van Atteveldt, W., Van der Velden, M. A. and Boukes, M.: 2021, The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms, *Communication Methods and Measures* **15**(2), 121–140.

Wang, X., Yang, Y., Deng, Y. and He, K.: 2021, Adversarial training with fast gradient projection method against synonym substitution based text attacks, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 13997–14005.

Webersinke, N., Kraus, M., Bingler, J. A. and Leippold, M.: 2021, Climatebert: A pretrained language model for climate-related text, *arXiv preprint arXiv:2110.12010* .

Yang, Y., Uy, M. C. S. and Huang, A.: 2020, Finbert: A pretrained language model for financial communications, *arxiv preprint arxiv:2006.08097*.

Zhu, Y., Hoepner, A. G., Moore, T. K. and Urquhart, A.: 2022, Sentiment analysis methods: Survey and evaluation, *Available at SSRN 4191581* .

Supplementary Material

A Prompting GPT-3

A.1 Rephrasing sentences to attack keyword-based approaches

```
#for sent, idx in  zip(df[((df["KW"] == ’negative’))&(df["Sentiment"]==’negative’)][’Headline’],df[((df["KW"] == ’negative’))&(df["Sentiment"]==’negative’)].index) :
for sent, idx in  zip(df[((df["KW"] == ’negative’))&(df["Sentiment"]==’negative’)&(df["KWx"]==’negative’)][’Headline’],df[((df["KW"] == ’negative’))&(df["Sentiment"]==’negative’)&(df["KWx"]==’negative’)].index) :
  negativewords = toneCntWneg(lmdict,sent)[2]
  allWords = []
  for nw in negativewords:
    response = openai.Completion.create(
          model="text-davinci-003",
          prompt= ["\"\nQuestion: Generate 30 single-worded synonyms to the following word and give it as a comma-seperated python list The synoyms should fit the context of the sentence: " + sent + ’ Avoid special characters. Word: ’ + nw  + " \nAnswer:"],
          temperature= 0.6,
          max_tokens= 256,
          top_p=1,
          frequency_penalty=0,
          presence_penalty=0
     )
    new_list = [item.strip() for item in response[’choices’][0][’text’].split(",")]
    candidates = []
    for word in new_list:
            candidates.append(check(word, lm_neg))
    newList = list(filter(None, candidates))
  allWords.append(newList)
  response = openai.Completion.create(
       model="text-davinci-003",
       prompt= ["\"\nQuestion: Rephrase the sentence" + str(sent) + "by substituting the word " + str(negativewords) + " with the best-fitting finance-related synonym from the following lists, so that the context, meaning, and grammar is still correct. List:" + str(allWords) + ". Do not change other negative sounding words. " + "\nAnswer:"],
       temperature=0.6,
       max_tokens=256,
       top_p=1,
       frequency_penalty=0,
       presence_penalty=0
    )
  df["Sent Syn x"][idx] = response["choices"][0]["text"]
  print(allWords)
  print(lex(df["Sent Syn x"][idx] ))
```

Listing 1: Generating adversarial headlines.

A.2 Keyword generation

In Listing 2, I list the code used in order to generate a keyword list that can be used for a larger-scale analysis, see Listing 3. The whole generated keyword list can be found here, in json-format.

```
# create a dictionary of all the negative words that are part of the LM dictionary
dicts = {}
keys = df_nw[’NegWord’].to_list()

k = 0
for i in range(len(keys)):
  if k < 0:
    k=k+1
    print(k)
    continue
  else:
    k=k+1
    print(k)
    response = openai.Completion.create(
      model="text-davinci-003",
      prompt= ["\"\nQuestion: Generate 100 single-worded synonyms
               for the financial context to the following word.
               Word: " + str(keys[i]) + " \nAnswer:"],
      temperature=0.6,
      max_tokens=256,
      top_p=1,
      frequency_penalty=0,
      presence_penalty=0
    )
    new_list = [item.strip() for item in response[’choices’][0][’text’].split(",")]
    candidates = []
    for word in new_list:
      candidates.append(check(word, lm_neg))
    dicts[keys[i]] = candidates
```

Listing 2: Generating synonyms with GPT-3 that are not part of the LM dictionary.

A.3 Reformulating sentences

```
for sent, idx in  zip(subsample[’Headline’], subsample.index):
  negativewords = toneCntWneg(lmdict,sent)[2]
  synonymlist = []
# identify the negative words in a given sentence and grab the
# synonyms (previously generated with GPT-3)
  for word in negativewords:
    synonymlist.append(df_syn[df_syn["LM Word"] == word]["Synonyms"].
   to_list())
 # access GPT-3
  response = openai.Completion.create(
       model="text-davinci-003",
       prompt= ["\"\nQuestion: Rephrase the sentence" + str(sent)
       + "by substituting the word " + str(negativewords)
       + " with the best-fitting finance-related synonym from
       the following lists, so that the context, meaning,
       and grammar is still correct. List:" + str(synonymlist) +".
       Do not change other negative-sounding words. " + "\nAnswer:"],
       temperature=0,
       max_tokens=256,
       top_p=1,
       frequency_penalty=1,
       presence_penalty=1
     )
  df["New Sentence"][idx] = response["choices"][0]["text"]
```

Listing 3: Prompting GPT-3.

B Example sentences and adversarial attacks

Table B.1: Human-annotated examples from the financial phrase bank Malo et al. (2013).

Sentence: The business to be divested generates consolidated net sales of EUR 60 million annually and currently has some 640 employees.
Label: **neutral**
Sentence: Svyturys-Utenos Alus, which is controlled by the Nordic group Baltic Beverages Holding (BBH), posted a 6.1 percent growth in beer sales for January-September to 101.99 million liters.
Label: **positive**
Sentence: The Department Store Division's sales fell by 8.6% to EUR 140.2 mn.
Label: **negative**
Sentence: Production capacity will rise gradually from 170,000 tonnes to 215,000 tonnes.
Label: **positive**
Sentence: Rautalinko was responsible also for Mobility Services, and his job in this division will be continued by Marek Hintze.
Label: **neutral**
Sentence: Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008.
Label: **positive**
Sentence: The changes will take effect on 1 January 2010, and they are not estimated to have an impact on the number of employees.
Label: **neutral**

Table B.2: Attacked examples from financial phrase bank.

The table reports some representative sentences from the financial phrase bank. I selected only sentences that humans and the keyword-based approach labeled as negative. By KWx, I denote the label given to the attacked sentence by the keyword-based approach. By FB and FBx, I denote the labels given by FinBERT to the original and the attacked sentence, respectively.

Sentence	Attacked Sentence	KWx	FB	FBx
Jan. 6 – Ford is struggling in the face of slowing truck and SUV sales and a surfeit of up-to-date , gotta-have cars .	Ford is struggling in the face of decelerating truck and SUV sales and a surfeit of up-to-date , gotta-have cars.	neutral	negative	negative
Asian traffic declined by 3.4 per cent .	Asian traffic fell by 3.4 per cent.	neutral	negative	negative
Myllykoski , with one paper plant in Finland , one in the US and three in Germany , had revenues of EUR286m in the first half of 2010 and an operating loss of EUR12m , Reuters said .	Myllykoski, with one paper plant in Finland, one in the US and three in Germany, had revenues of EUR286m in the first half of 2010 and an operating decrease of EUR12m, Reuters said.	neutral	negative	negative
2009 3 February 2010 - Finland-based steel maker Rautaruukki Oyj (HEL : RTRKS) , or Ruukki , said today it slipped to a larger-than-expected pretax loss of EUR46m in the fourth quarter of 2009 from a year-earlier profit of EUR45m .	Ruukki said today it reversed to a larger-than-expected pretax liability of EUR46m in the fourth quarter of 2009 from a year-earlier profit of EUR45m.	neutral	negative	negative
In this case , the effect would be negative in Finland .	In this case , the effect would be ruinous in Finland .	neutral	neutral	negative
Nokia will certainly disagree with Qualcomm 's views on the patent situation .	Nokia will certainly gainsay Qualcomm's views on the patent situation.	neutral	negative	negative
The number of bodily injury cases quadrupled in 2000-2006 .	The number of affliction cases quadrupled in 2000-2006.	neutral	negative	negative
In January-June 2010 , diluted loss per share stood at EUR0 .3 versus EUR0 .1 in the first half of 2009 .	In January-June 2010 , diluted reduction per share stood at EUR0 .3 versus EUR0 .1 in the first half of 2009 .	neutral	negative	negative
Operating loss totalled EUR 5.2 mn , compared to a loss of EUR 3.4 mn in the corresponding period in 2008-2009 .	Operating drain totalled EUR 5.2 mn, compared to an abatement of EUR 3.4 mn in the corresponding period in 2008-2009.	neutral	negative	negative
Operating loss totaled EUR 0.3 mn compared to a profit of EUR 2.2 mn in the corresponding period in 2007 .	Operating reduction totaled EUR 0.3 mn compared to a profit of EUR 2.2 mn in the corresponding period in 2007.	neutral	negative	negative

Table B.3: Attacked examples from earning calls.
The table reports some representative sentences from earning calls for which the keyword-based approach provides a negative sentiment. The original sentences are given under *a)* and the attacked sentences under *b)*. Also reported are the verdicts for the original and new sentences for the keyword-based approach, FinBERT, and the human intuition of this paper's author.

Sentence	Keyword	FinBERT	human
1a) Obviously, we had also the COVID-19 pandemic and certain shortage of labor.	**negative**	**negative**	**negative**
1b) Obviously, we had also the COVID-19 pandemic and certain scarcity of labor.	**neutral**	**negative**	**negative**
2a) And despite significant challenges imposed by a sharp rise in inflation levels, high nat cat losses and ongoing claims from the COVID-19 pandemic, we delivered an excellent result of EUR 2.9 billion.	**negative**	**positive**	**positive**
2b) And despite significant predicaments imposed by a sharp rise in inflation levels, high nat cat reductions and ongoing remunerations from the COVID-19 pandemic, we delivered an excellent result of EUR 2.9 billion.	**positive**	**positive**	**positive**
3a) This is due to mainly three reasons: significantly higher price-driven sales comparing to unchanged absolute EBITDA, resulting in a mathematically lower margin; earnings significantly burdened by higher feedstock prices; and longer-term sales contracts yet presented to fully pass-through prices, reflecting the typical nature of a specialty business.	**negative**	**negative**	**negative**
3b) This is due to mainly three reasons: significantly higher price-driven sales comparing to unchanged absolute EBITDA, resulting in a mathematically lower margin; earnings significantly charged by higher feedstock prices; and longer-term sales contracts yet presented to fully pass-through prices, reflecting the typical nature of a specialty business.	**neutral**	**negative**	**negative**

Swiss Finance Institute

Swiss Finance Institute (SFI) is the national center for fundamental research, doctoral training, knowledge exchange, and continuing education in the fields of banking and finance. SFI's mission is to grow knowledge capital for the Swiss financial marketplace. Created in 2006 as a public–private partnership, SFI is a common initiative of the Swiss finance industry, leading Swiss universities, and the Swiss Confederation.

swiss:finance:institute

c/o University of Geneva, Bd. Du Pont d'Arve 42, CH-1211 Geneva 4
T +41 22 379 84 71, rps@sfi.ch, www.sfi.ch