Finance Research Letters 55 (2023) 103957

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

Sentiment spin: Attacking financial sentiment with GPT-3

Markus Leippold

Department of Banking and Finance, University of Zurich, Switzerland
Swiss Finance Institute (SFI), Plattenstrasse 14, 8032, Switzerland

ARTICLE INFO

JEL classification:
G2
G38
C8
M48

Keywords:
sentiment analysis in financial markets
Keyword-based approach
FinBERT
GPT-3

ABSTRACT

In this study, we explore the susceptibility of financial sentiment analysis to adversarial attacks that manipulate financial texts. With the rise of AI readership in the financial sector, companies are adapting their language and disclosures to fit AI processing better, leading to concerns about the potential for manipulation. In the finance literature, keyword-based methods, such as dictionaries, are still widely used for financial sentiment analysis due to their perceived transparency. However, our research demonstrates the vulnerability of keyword-based approaches by successfully generating adversarial attacks using the sophisticated transformer model, GPT-3. With a success rate of nearly 99% for negative sentences in the Financial Phrase Bank, a widely used database for financial sentiment analysis, we highlight the importance of incorporating robust methods, such as context-aware approaches such as BERT, in financial sentiment analysis.

1. Introduction

This paper explores whether financial texts can be manipulated to deceive machine readers in their sentiment analysis. As highlighted by Cao et al. (2022), it becomes crucial to understand the potential consequences of the increasing artificial intelligence (AI) readership on corporate filings. As firms tailor their financial disclosures to cater to machine processing, it raises concerns about the accuracy and reliability of sentiment and other textual analyses. The finance literature predominantly uses keyword-based approaches, such as the (Loughran and McDonald, 2011) dictionary, to sentiment analysis, which are rule-based and rely on the presence or absence of certain words or phrases to determine overall sentiment.[1] However, recent research suggests that these approaches may not always provide reliable or accurate results, particularly when applied to more complex or nuanced texts.[2] The paper does not aim to debate the effectiveness of keyword-based versus more advanced sentiment classification methods but instead examines the vulnerability of sentiment classification approaches to adversarial attacks.

Previous research on adversarial attacks in natural language processing (NLP) primarily used rule-based methods to create negative examples by replacing words with synonyms. However, manipulations with these methods can lead to unnatural and confusing changes in the text that humans can easily detect. Recent advances in NLP have led to renewed interest in adversarial attacks in the field, and various methods have been proposed.[3] However, studies have shown that even advanced methods do not always preserve the semantics of the original text, and between 96% and 99% of analyzed attacks do not preserve semantics.[4]

E-mail address: markus.leippold@bf.uzh.ch.

[1] As of January 2023, the paper of Loughran and McDonald (2011) has been cited more than 4276 times by researchers. Moreover, their word list has been adopted for the WRDS SEC Sentiment Data.

[2] See, e.g., Boukes et al. (2020), Hartmann et al. (2022), Van Atteveldt et al. (2021) and Webersinke et al. (2022).

[3] See, e.g., Papernot et al. (2016), Alzantot et al. (2018), Jin et al. (2020), Garg and Ramakrishnan (2020) and Wang et al. (2021).

[4] See Morris et al. (2020) and Hauser et al. (2021).

https://doi.org/10.1016/j.frl.2023.103957
Received 4 March 2023; Received in revised form 9 April 2023; Accepted 28 April 2023
Available online 5 May 2023

Therefore, to construct adversarial attacks, the article relies on the eloquence of GPT-3, a large language model for generating contextually and semantically correct text.

The study highlights the importance of addressing the potential vulnerabilities of algorithmic sentiment classification approaches to adversarial attacks, particularly as the algorithmic readership of financial texts increases. The findings suggest that GPT-3-generated adversarial attacks can deceive algorithms when solving sentiment analysis tasks. The article emphasizes the need for developing more robust and reliable sentiment analysis methods to address these vulnerabilities.

2. NLP methods

I will briefly give an overview of the methods used in this study. The (Henry, 2008) Financial Dictionary and the (Loughran and McDonald, 2011) dictionary (LM) are the most prominent financial dictionaries. The former dictionary has only a limited number of words and low coverage.[5] In contrast, the LM dictionary is a sentiment word list compiled from firms' annual reports and includes 354 positive, 2355 negative, 297 uncertainty, 904 litigious, 19 strong modal, 27 weak modal, and 184 constraining words. LM is the most widely used finance domain lexicon we know. Therefore, we focus on the LM in the subsequent analysis.

Recent advancements in architecture and training methods have significantly improved the way deep learning models can learn general language to perform various downstream tasks. The transformer architecture of BERT (Devlin et al., 2018) and other deep neural networks allow a truly bi-directional contextual relationship to be learned. These models are often trained on a large corpus of general text and can 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).[6] The transformer architecture allows FinBERT to process input text more sophisticatedly by selectively focusing on different input parts. 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.

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, summarization, 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 various applications. I will use GPT-3 for the text generation required for adversarial attacks. However, in the Appendix, I present some interesting results on sentiment classification (which is not considered a tough task in NLP).

3. Data

For my experiments, I use an open-sourced database of human-annotated sentiments, the well-known Financial Phrase Bank developed in Malo et al. (2013), which contains 4837 English news headlines of companies listed on the OMX Helsinki exchange.[7] The total number of sentences corresponds to the instances of an inner-annotator agreement larger than 50%. From these sentences, there are 2264 sentences with a 100% agreement. My experiment will use the full database with 4846 sentences, consisting of 604 negative, 2879 neutral, and 1363 positive sentences. Examples of the sentences and their labels from human annotations are given in Table C.3. 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. I will use unlabeled data from earning calls for the second part of my experiments.

4. Manipulating sentiment with GPT-3

By using GPT-3, which is one of the most advanced models in conversational AI, 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.[8]

[5] Which therefore also makes it more vulnerable to adversarial attacks.

[6] 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).

[7] 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.

[8] 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 GPT-3. For more recent advances on LLMs and their potential use, see, e.g., Ashraf Vaghefi et al. (2023) and Kraus et al. (2023).

4.1. Strategy 1: Prompting GPT-3

In Strategy 1, I will follow a very direct approach in prompting GPT-3.[9] In particular, I will apply the code in Listing 1, given in Appendix B, 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.

4.2. 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 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 B, together with the list of generated synonyms for all the negative words in the LM dictionary.

5. 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.

5.1. Manipulating sentiment in the Financial Phrase Bank

In the data, we can find 256 sentences that 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.

Fig. 1 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 attack’s success rate is at 99% for the keyword-based approach, while it is only at 7% for FinBERT. Table C.4 in Appendix C 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’, 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.

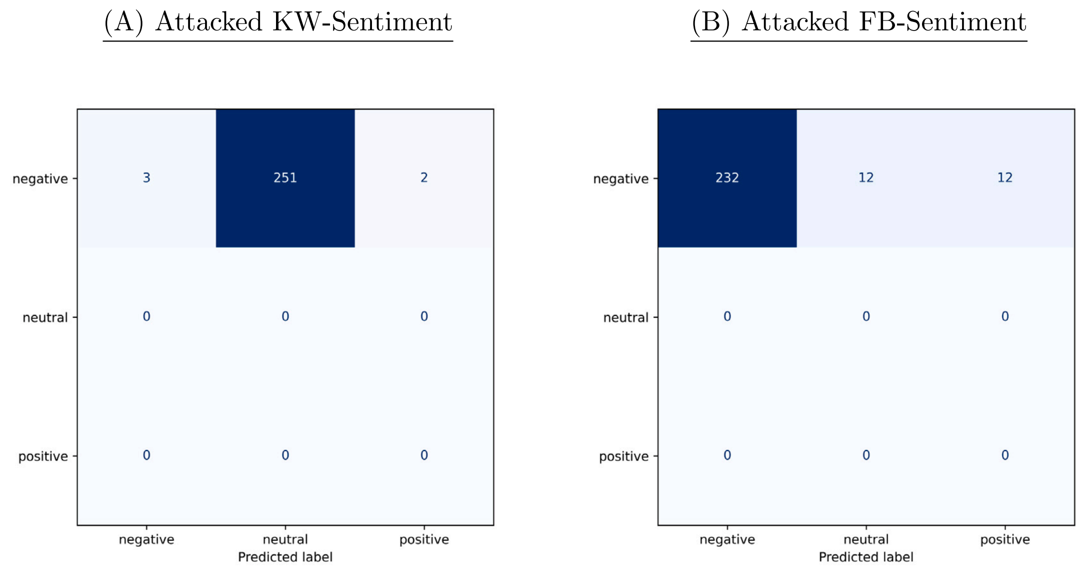
5.2. Manipulating sentiment in earning calls

For my second experiment, I use sentences from the presentation sections of earning calls in 2022. In Fig. 2, Panel A illustrates 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.

Fig. 3 shows the distributions of sentiment scores computed using the keyword-based approach for 35,000 sentences from call transcripts in 2022.[10] The sample size of sentences with negative sentiment obtained by the keyword-based approach is 3902. 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 sentences in the presentation section for a given company. Panel B of Fig. 3 shows the distribution of sentiment scores when computed with the FinBERT model, again for the original sentences and the

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

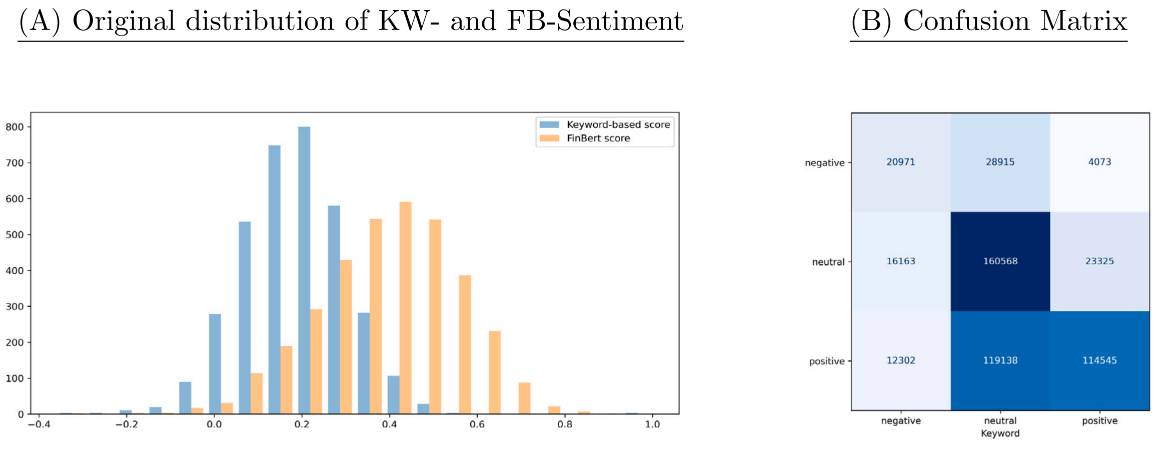
[10] I have reduced the number of sentences attacked by GPT-3 to 3902 due to computational constraints. However, the results hold without loss of generality.


(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**

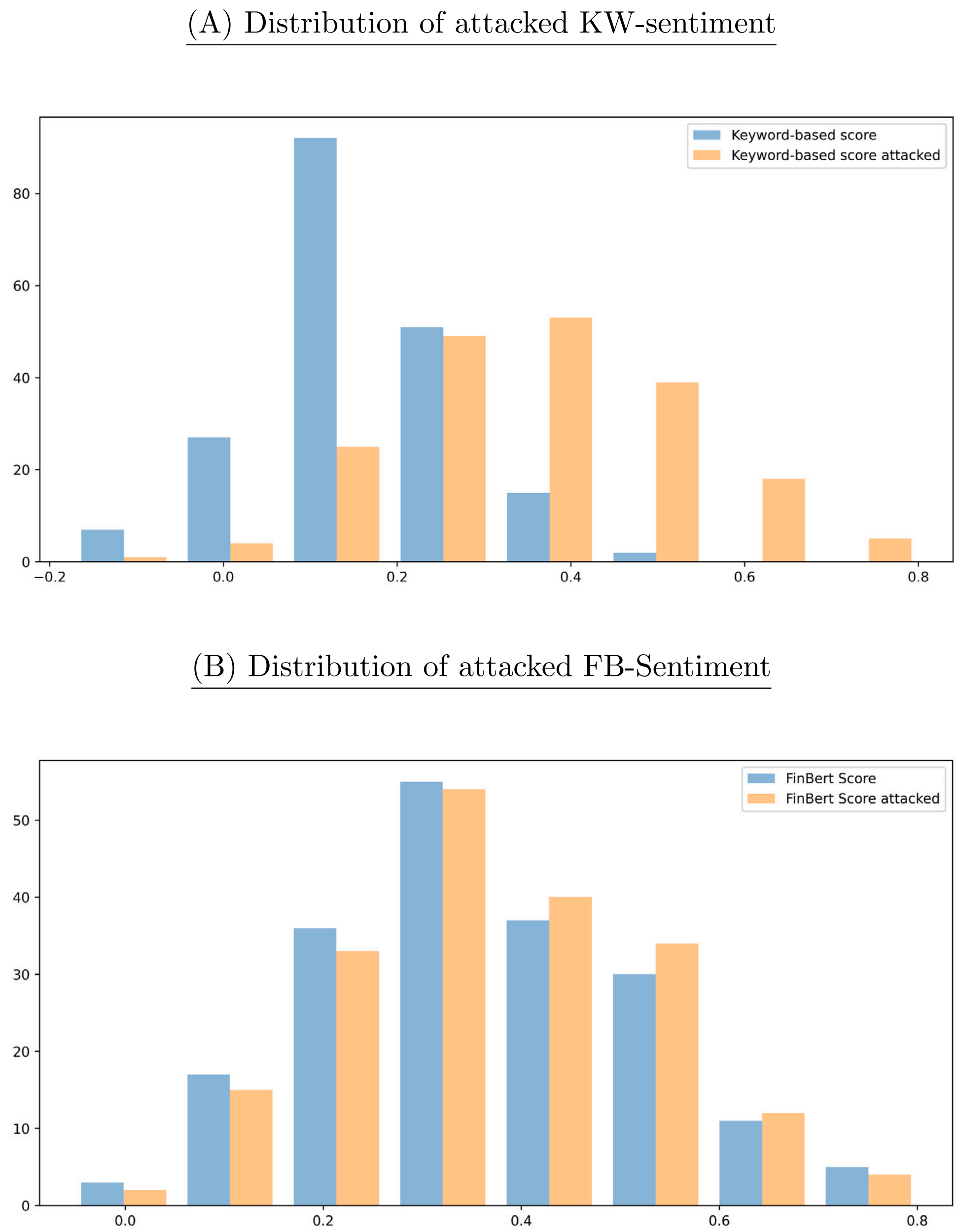
Fig. 1. The figure shows the confusion matrix for the attacked sentences with negative sentiment. The sample consists only of sentences that originally have been correctly identified by the keyword-based approach with a negative sentiment. The table below reports the different performance measures. In 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%.


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

(B) Confusion Matrix

Fig. 2. 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 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 C.5 in Appendix C, together with the verdicts of the keyword-based approach and FinBERT and my own (human) assessment.



(A) Distribution of attacked KW-sentiment

(B) Distribution of attacked FB-Sentiment

Fig. 3. 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 3902. 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.

6. Conclusion

In conclusion, I have demonstrated the vulnerability of keyword-based sentiment analysis methods to adversarial attacks using GPT-3, even when leveraging widely-used dictionaries such as the one by Loughran and McDonald (2011). My experiments have shown that an LLM like GPT-3 can effectively change the sentiment of sentences without significantly altering their semantic quality, which exposes the shortcomings of keyword-based approaches. On the other hand, more advanced models like FinBERT display higher resilience to adversarial manipulation, indicating their robustness and superiority in detecting sentiment.

While this paper purposely employs a relatively simple approach to generate adversarial attacks, it highlights the capabilities of modern NLP methods, opening up possibilities for developing more potent and sophisticated adversarial techniques. Given the increasing reliance on AI-powered information processing amidst the ongoing information overload, exploring adversarial attacks and developing robust NLP models becomes a crucial area of research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.frl.2023.103957.

References

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.-J., Srivastava, M., 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.

Ashraf Vaghefi, S., Wang, Q., Muccione, V., Ni, J., Kraus, M., Bingler, J., Schimanski, T., Colesanti-Senni, C., Webersinke, N., Huggel, C., et al., 2023. Chatclimate: Grounding conversational AI in climate science. pp. arXiv–2304, arXiv e-prints.

Boukes, M., Van de Velde, B., Araujo, T., Vliegenthart, R., 2020. What's the tone? Easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. Commun. Methods Meas. 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. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Cao, S., Jiang, W., Yang, B., Zhang, A.L., 2022. How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI. SSRN Working Paper.

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

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

Hartmann, J., Heitmann, M., Siebert, C., Schamp, C., 2022. More than a feeling: Accuracy and application of sentiment analysis. Int. J. Res. Mark. 40 (1).

Hauser, J., Meng, Z., Pascual, D., 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? J. Bus. Commun. (1973) 45 (4), 363–407.

Jin, D., Jin, Z., Zhou, J.T., Szolovits, P., 2020. Is Bert really robust? A strong baseline for natural language attack on text classification and entailment. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. pp. 8018–8025.

Kraus, M., Bingler, J.A., Leippold, M., Schimanski, T., Senni, C.C., Stammbach, D., Vaghefi, S.A., Webersinke, N., 2023. Enhancing large language models with climate resources. arXiv preprint arXiv:2304.00116.

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

Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J. Finance 66 (1), 35–65.

Malo, P., Sinha, A., Takala, P., Korhonen, P.J., 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.

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

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

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

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

Webersinke, N., Kraus, M., Bingler, J., Leippold, M., 2022. Climatebert: A pretrained language model for climate-related text. In: Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges.

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