An Adversarial Attack Approach on Financial LLMs Driven by Embedding-Similarity Optimization

Master's Thesis

Submitted in partial fulfillment of the requirements for the degree of Master of Arts in Economics

Author

Aysun Can Turetken

20-741-195

aysun.cantueretken@uzh.ch

Supervisor

Prof. Dr. Markus Leippold

Professor of Financial Engineering

Department of Banking and Finance

University of Zurich

18.07.2024

Abstract

Adversarial attacks on financial sentiment analysis models are a critical area of research within NLP. We introduce a novel white-box attack method that leverages a pre-trained general-purpose language model to generate high-quality and human-imperceptible attacks. Unlike existing methods that rely on training specialized adversarial models or computationally-intensive gradient optimization routines, our approach employs carefully-designed instructions and a novel embedding-similarity function to maintain semantic integrity while producing linguistically rich adversarial samples without requiring ground-truth sentiment labels. Our results, obtained from attacking FinBERT and FinGPT models across three public datasets, demonstrate significant performance degradation for the two models, with MAE values reaching 0.67 and 0.48, respectively.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Adversarial attacks on language models are becoming a significant area of research in Natural Language Processing (NLP), particularly in the context of Large Language Models (LLMs). By exploiting vulnerabilities in language models trained for specific or general tasks, attackers can inflict financial losses (Xie and Wang, 2022), disrupt business operations, breach data privacy (Usynin et al., 2021), spread disinformation, and obtain harmful or sensitive information (Zou et al., 2023). Therefore, understanding the failure modes of these models is crucial for identifying and mitigating vulnerabilities in NLP systems.

Financial sentiment analysis models, which interpret complex and often subtle linguistic cues from vast amounts of unstructured data to influence sentiment-driven financial decisions, embody a critical application of NLP where the stakes are particularly high. In the financial sector, the development of models that are resilient to adversarial attacks becomes increasingly crucial. This necessity stems from the growing reliance on automated processes within the industry, which demands not only accurate but also secure NLP applications to maintain trust and efficacy in numerous decision-making processes.

In finance, these processes are mostly driven by classification tasks, such as fraudulent transaction identification, credit scoring, risk classification, named entity recognition and sentiment classification, many of which are used to interpret and monitor the financial markets. The specialized and intricate language used in finance has so far prompted structured or controlled adversarial attack approaches for these classification tasks in an attempt to maintain the semantics of the input financial texts as much as possible. Many general-purpose character-level techniques are unsuitable for this reason as they pose the risk of generating semantically or factually incorrect text. Word-level attacks do enforce semantic similarity by carefully choosing the words to be replaced (e.g., using synonyms) or removed, which is usually achieved using gradient-based optimization techniques. This is why, they have been favored over other approaches so far in the financial domain (Xie and Wang, 2022; Leippold, 2023). However, these methods only encompass a restricted subset of all conceivable semantically-consistent attack strategies, highlighting the need for more effective alternatives.

Sentence-level attacks attempt to address this limitation by allowing for broader perturbations such as syntactic changes, sentence paraphrasings or addition of suffix phrases (Iyyer et al., 2018; Wang, Pei, et al., 2020; Zou et al., 2023). However, most of these methods train a purpose-built adversary model, which is time-consuming and costly to do, and may lack the advanced language articulation capabilities necessary for paraphrasing complex and nuanced financial texts effectively.

In this thesis, in contrast to many current adversarial attack methodologies, we introduce a simple but effective white-box approach that obviates the need for training a specialised adversarial model or dependence on gradients and complex optimization schemes, which are computationally expensive and necessitate a training dataset. Instead, our method utilizes an existing pre-trained general-purpose language model (GPT-4o) to generate high-quality, human-imperceptible attacks. By employing meticulously designed instructions for generating and filtering adversarial samples, and incorporating a novel embedding-similarity-based loss function, we enforce semantic integrity while producing syntactically diverse and linguistically rich samples. Our approach does not rely on target (ground-truth) classification labels, and instead, aims to alter the response of the classification model in their absence. This is particularly relevant in finance as such labels are frequently unavailable to the attackers and the public in many practical scenarios.

We demonstrate the effectiveness of our approach for the financial sentiment classification task by conducting adversarial attacks on two financial models, FinBERT and FinGPT, using three public datasets. One of these datasets is employed for fine-tuning the models, while the other two remain unseen by them. Our results indicate that the responses of FinBERT and FinGPT are respectively altered in approximately 50% and 40% of the samples from the unseen datasets, and in 20% and

30% of the samples from the fine-tuning dataset. Despite not relying on ground truth sentiment labels, our attack results in more than 20% drop in accuracy for both models on the samples they were fine-tuned on.

In the subsequent chapters, we begin with an overview of recent adversarial attack techniques on classification LLMs that are relevant to our work. In Chapter 3, we detail our approach, and in Chapter 4, we describe our experimental setup involving the datasets and LLM models utilized. Chapter 5 presents our results and findings, including the identified failure modes of the two financial models attacked. Finally, we summarize our conclusions in Chapter 6 and discuss potential directions for future research.

Chapter 2

Literature Review

In this section, we give a general overview of recent textual adversarial attacks on NLP classification models including both general purpose and finance-specific ones, the latter of which has a rather limited number of examples. In the following, we first start with local approaches involving character-level attacks and then gradually move on to more global ones including word-and sentence-level attacks.

2.1 Character-Level Attacks

Character-level attacks involve perturbations of certain characters in the original text to trick a text classifier into giving an incorrect output. This is typically done by first identifying important characters, tokens or words and then altering them using predetermined rules, which involve character-level substitutions, deletions, insertions and swapping.

A notable example of a character level perturbation method is HotFlip (Ebrahimi et al., 2017), which functions under white-box settings. It uses a gradient-based optimization technique on the one-hot representation of the input text to identify the optimal character changes. Conversely, Charmer (Abad Rocamora et al., 2024) operates under black-box settings by only observing the output predictions of a text classifier. It first finds the most impactful character positions to modify by maximizing a loss function between the target labels and the model predictions.

In contrast to these approaches that focus directly on characters for perturbations, DeepWord-Bug (Gao et al., 2018) first identifies important words using a rule-based function that evaluates the model multiple times for each word. It then applies pre-defined transformations such as character swaps and substitutions to mislead the classifier. Similarly, TextBugger (Li, Ji, et al., 2018) first finds important words, in both white-box (using the Jacobian matrix of the classifier) and black-box (using model predictions) scenarios, and then chooses an optimal perturbation strategy that preserves semantic similarity between the original and the perturbed words.

Although these techniques achieve high performance on several textual datasets, some of the adversarial samples they generate look unnatural and are human perceptible as they do not resemble human-generated typos.

2.2 Word-Level Attacks

Unlike character-level attacks, which involve altering individual characters as the atomic building blocks of language, word-level attacks involve manipulations through word insertions, deletions, or replacements. A pioneering work in this category utilizes the targeted model's gradients (Papernot et al., 2016) to find the optimal world-level attack. More specifically, it computes the Jacobian matrix to generate adversarial samples that deceive the model while minimizing the number of word replacements to the original text. However, the attacks generated by this approach are often human perceptible due to the semantic inconsistencies it introduces.

Similarly, the method proposed in (Samanta and Mehta, 2017) adopts a gradient-based approach, called Fast Gradient Sign Method (FGSM), to rank the words according to their contributions (i.e., importance) towards the predicted class label. It then iterates over the ranked words and performs perturbations on each until the model's prediction changes. Despite employing various heuristics to maintain the semantic meaning of the sentences, the resulting attacks remain limited in ensuring semantic consistency.

Among the black-box techniques, a notable approach (Alzantot et al., 2018) leverages a genetic algorithm that evolves a population of perturbed examples across multiple generations while incorporating Euclidean distance and GloVe word-level embeddings (Pennington et al., 2014) to maintain semantic integrity. Likewise, TextFooler (Jin et al., 2020) adopts a rule-based black-box strategy, utilizing importance-based word substitution. It utilizes word embeddings to extract synonyms for word replacement and the Universal Sentence Encoder (USE) (Cer et al., 2018), which provides sentence-level embeddings to assess the semantic similarity between the original and modified sentences.

The introduction of pre-trained deep architectures for language, such as BERT, marked a shift in NLP research from rule-based to more automated methods. BERT-Attack (Li, Ma, et al., 2020) and BAE (Garg and Ramakrishnan, 2020) were among the initial studies to exploit BERT's capabilities to develop word-level attacks. These works leverage BERT's context-sensitive prediction abilities to craft adversarial samples that are semantically consistent with the original samples. More specifically, BERT-Attack focuses on identifying important key words and employs BERT's Masked Language Model to generate suitable word substitutions. Analogously, BAE adopts a comparable strategy but extends its perturbations to include both word replacements and insertions.

Another recent approach that rely on BERT to preserve semantic consistency is GBDA (Guo et al., 2021). However, in contrast to BERT-Attack and BAE, which generate a single adversarial attack in a black-box setting, GBDA generates a distribution of adversarial samples through a gradient-based optimization technique. In estimating this optimal adversarial distribution, the method minimizes a differentiable triple-loss function including attacked model's adversarial loss, a semantic similarity loss based on BERT embeddings and a language fluency loss based on perplexity. However, despite the use of the fluency and semantic similarity terms, the generated attacks still exhibit distorted semantics as shown by their examples.

In the financial domain, existing approaches has so far concentrated mostly on world-level attacks. This is partly because many character-level methods cannot be readily applied to the financial corpora since they run the risk of altering the financial sentiment of the original samples. A straightforward example to this is altering the numerical values or adding (or removing) a negative sign before a number.

For this reason, the specialized language used in finance has led to adversarial attack approaches that employ a more structured or controlled methodologies. For instance, the world-level approach of (Xie and Wang, 2022) formulates the adversarial attack on directional stock forecast models as a combinatorial optimization problem defined over both the textual and temporal dimensions. It solves this problem by using a sophisticated gradient based optimization algorithm that finds the optimal tweets, words and their replacements (or deletions). As with several other earlier approaches, it uses GLoVe embeddings (Pennington et al., 2014) to ensure the semantic similarity for word replacements.

The method of (Leippold, 2023) employs more direct optimization-free approach using GPT-3 (Brown et al., 2020) to craft word-level adversarial samples. It reveals the capabilities of large language models in generating high-quality adversarial attacks as well as the vulnerabilities of keyword-based approaches used in financial sentiment analysis. The method first identifies the negative words in the input financial sentences by using a negative subset of the financial words dictionary (LM) introduced in (Loughran and Mcdonald, 2011). It then generates a list of synonyms for these words using the GPT-3 model while adhering to the financial context of the input sentences, and subsequently filters out those synonyms that also exist in the negative LM subset. Finally, it prompts GPT-3 again to replace the original negative words with the remaining synonyms while preserving the context and meaning with correct grammar.

Word-level attacks often produce more natural and semantically-preserving results than character-level ones, particularly with recent advancements in the LLM domain. However, within the space of all possible semantically-consistent attacks, these methods still represent only a small subspace, leaving considerable room for improvement. Sentence-level attacks, described in the following section, attempt to address this gap.

2.3 Sentence-Level Attacks

The primary feature of sentence-level attacks is that they apply at a broader level over the input text compared to character and word-level attacks. This includes comprehensive modifications such as controlled structural alterations (syntactic changes), sentence paraphrases and addition of new phrases to the beginning or end of a sample to confuse the classifier while maintaining its target category label (i.e., ground truth class).

A prominent method explored in the literature involves generating sentence paraphrases. Syntactically controlled paraphrase networks (SCPNs) represent one such approach (Iyyer et al., 2018). It trains a neural encoder-decoder network that generates rephrasings of the original sentence while adhering to predetermined syntactic structures. The output of this network are then post-processed to filter out semantically invalid paraphrases using n-gram overlap and paraphrastic similarity obtained from a pre-trained sentence embedding model (Wieting and Gimpel, 2018).

Analogous to SCPNs, T3 (Wang, Pei, et al., 2020) framework also utilizes an encoder-decoder architecture but introduces a tree-based design, which leverages the structural and contextual elements of text. The tree-based encoder embeds discrete text into a continuous space which allows to search for adversarial perturbations in word or sentence level. The tree-based decoder is then used to map perturbed embeddings back to adversarial sentences while seeking to preserve both the semantic content and syntactic structures of the original input.

Similarly, the black-box approach of (Zhao et al., 2018) relies on a training an adversary model that aims to keep the semantic meaning intact. However, instead of an encoder-decoder architecture, it uses Generative Adversarial Networks (GANs) to create adversarial samples for various tasks including textual entailment. It achieves this by training a generator and a critic sub-network, where the generator creates a text sample from an input noise with a certain probability distribution and the critic attempts to distinguish it from real samples. It also trains an inverter network which can map a given text sample back to the noise space. At inference time, the method first runs an input sample through the inverter, perturbs the resulting noise vector and passes it to the generator to create an adversarial sample that is semantically close to the original input. This approach has been used in the finance domain to perform fine-tuning of existing models on the generated adversarial samples to improve their classification performance for stock sentiment prediction (Wang and Gan, 2023).

The literature on sentence-level attacks in classification tasks is limited, whereas such attacks are more prevalent in downstream NLP tasks. For instance, the addition of irrelevant sentences has been explored as an attack method in question answering (Jia and Liang, 2017), and white-box gradient-based methods (Cheng et al., 2019) have been applied in neural machine translation. These examples illustrate the diversity of approaches in sentence-level attacks across various NLP applications beyond the classification-related ones.

Contrary to many earlier sentence-level techniques, we propose a simple white-box adversarial attack approach, which does not require training a task-or domain-specific adversary model, and instead, leverages existing pre-trained general-purpose LLMs for generating high-quality and human-indistinguishable attacks. Through the use of carefully-crafted LLM instructions and a new embedding-similarity based loss function, we enforce semantic similarity while generating linguistically rich attack samples that are hard to correctly classify by specialized and relatively smaller financial models. The next chapter will formalize our approach and outline the detailed steps involved.

Chapter 3

Methodology

Our approach leverages existing powerful general-purpose LLMs that exhibit a proficient command of the English language to challenge smaller, specialized models in the domain of financial sentiment classification. Formally, for a given specialized model M to be attacked and an input text sample s, we first generate a set of N paraphrases $\{s^i\}_{i=0}^{N}$ using an advanced language model P, which is adept at comprehending and articulating a wide range of financial topics. This is achieved by prompting the advanced model P with a carefully crafted instruction I_1 to explicitly generate N paraphrases that are approximately the same length and preserve both the factual information and the financial sentiment of the original input. In Appendix B, we provide a few random input samples and their paraphrases to demonstrate the effectiveness of this approach.

We then compute the unit-vector embeddings of the original sample s and the paraphrased ones $\{s^i\}$ using both models. Let e_m and e_p be the embeddings computed from s using M and P respectively. Similarly, let e_m^i and e_p^i be the embeddings of the paraphrased sample s^i. Inspired by an embedding similarity-based search approach employed in a recent contamination detection method (Yang, Chiang, et al., 2023), we advocate that strong adversarial samples can be found by selecting the paraphrased sample that maximizes the ratio of the embedding similarity between e_p and e_p^i to the similarity between e_m and e_m^i:

$$s^* = \text{argmax}_{s^i} \left(\frac{\text{simil}(e_p^i, e_p)}{\text{simil}(e_m^i, e_m)} \right) \tag{3.1}$$

The primary rationale behind maximizing this function is to identify a paraphrased sample s_i, for which the two embeddings produced by M are sufficiently dissimilar, while those generated by P are similar. Among the N paraphrased samples, this process highlights the one that is most likely to alter (flip) the prediction of the classification model M, and hence yield a successful attack.

To ensure the validity of the optimal paraphrased sample s^*, we implement an additional precautionary step by using the model P a second time to verify if its sentiment matches that of the original input sample s. This is done by prompting P with a second instruction I_2, which explicitly asks it to predict the sentiment class of s and s^* independently, and verifies that the predictions are the same. If the sample s^* passes this verification step, it is considered the final adversarial sample. Otherwise, it is discarded, and the next sample that maximizes the embedding similarity ratio, as defined in Eq. (3.1), is selected.

In practice, we limit the number of verification iterations to three, as the most effective attacks usually occur within the first few attempts that maximize the embedding similarity ratio. If a sample that passes this filtering step can not be be found, we consider the attack unsuccessful and return the original input sample s. In our experiments, the percentage of such unsuccessful cases, however, remained below 7% for all the dataset and model pairs we evaluated.

Algorithm 1 summarizes our approach in the form of pseudo-code. To compute the similarity between the unit-vector embeddings, we use normalized cosine similarity, which produces non-negative values needed in ratio computation described above.

$$\text{simil}(e_i, e_j) = \frac{(1 + {e_i}^T \cdot e_j)}{2} \tag{3.2}$$

To generate the paraphrases, we use the following instruction I_1 in our algorithm:

I_1 : **Rephrase the following financial text in $\underline{N}$ different ways and present your response in a numbered list starting at 1. Be concise and make sure that each paraphrase has exactly the same factual information and financial sentiment: $\underline{s}$**

where the underlined parameter N is replaced with the desired number of paraphrases to be generated

Algorithm 1 Adversarial Attack Approach

Input: Financial text sample s, number of paraphrases N, instructions I_1 and I_2, attacking model P, attacked model M
Output: Optimal adversarial sample s^*
Step 1: Generate N paraphrased samples $\{s^i\}_{i=0}^N$ from s by prompting model P with I_1:
$\{s^i\}_{i=0}^N \leftarrow P(I_1(N, s))$
Step 2: Compute embeddings e_m and e_p for the input sample s using models M and P:
$e_m \leftarrow$ embedding of s using model M
$e_p \leftarrow$ embedding of s using model P
Step 3: Compute paraphrased sample embeddings e_m^i and e_p^i respectively using M and P:
for $i = 0$ to N **do**
    $e_m^i \leftarrow$ embedding of s^i using model M
    $e_p^i \leftarrow$ embedding of s^i using model P
end for
Step 4: Select the final adversarial sample s^* as the paraphrase that maximizes the embedding similarity ratio while maintaining the sentiment of the original input sample s:
$sent \leftarrow P(I_2(s))$
$iter \leftarrow 0$
repeat
    $s^* \leftarrow \text{argmax}_{s^i} \left(\frac{\text{simil}(e_p^i, e_p)}{\text{simil}(e_m^i, e_m)} \right)$
    $sent^* \leftarrow P(I_2(s^*))$
    if $sent^* \neq sent$ **then**
        Remove paraphrase s^* from the set $\{s^i\}$
        $s^* \leftarrow null$
    end if
    $iter \leftarrow iter + 1$
until $sent^* \neq sent$ and $\{s^i\} \neq \emptyset$ and $iter < 3$
Step 5: Return the optimal paraphrase as the final adversarial sample. If no such sample can be found, return the original input sample s:
if $s^* = null$ **then** $s^* \leftarrow s$
return s^*

with the model P. For our experiments, we set N to 25. The parameter s represents the input sample. We explicitly instruct the model to create concise paraphrases because, during preliminary experiments with multiple potential model candidates for P, we observed that they tend to generate lengthy and simplified explanatory statements in the absence of this directive. Such statements are easier for the attacked model M to interpret correctly, thereby enhancing its chances of accurately processing the sample text and counteracting our objective of attacking it. Enforcing conciseness ensures that the paraphrases are approximately the same length as the input text and eliminates redundant explanations.

In addition, we instruct the model P to maintain the same factual information and financial sentiment in the paraphrases it generates. This requirement stems from the nature of financial corpora, which are abundant in factual information such as key entities, market trends, monetary values, corporate announcements, policy changes, financial metrics, economic indicators, specific dates and fiscal years, all of which can influence the true sentiment classification of a sample. Thus, preserving factual accuracy is a critical constraint that we enforce through specific prompting.

Lastly, as described above, we use the second instruction I_2 to obtain a sentiment prediction from P independently for both the original input sample and a paraphrased one, and then check for their equivalence. We use the following instruction for this step:

I_2 : **Does the following text have a POSITIVE, NEGATIVE or NEUTRAL financial sentiment based only on the facts stated in the text? Output a single-word response representing the sentiment class:** $\underline{s}$

We specifically instruct the model to base its judgment solely on the facts stated in the input text. This approach is necessary because we observed that, due to its training on extensive public textual corpora, the advanced model P can recall details about certain financial events. Since these additional details are not present in the input text samples, they can skew its sentiment predictions.

In the following chapter, we describe our experimental setup, providing more details about the models we evaluated and the datasets used.

Chapter 4

Experimental Setup

In this chapter, we first provide a detailed discussion of the three datasets employed to demonstrate the performance of our adversarial attack approach. We then describe the two target LLM models subjected to the attacks and the advanced model used to generate the paraphrases introduced in the previous chapter.

4.1 Datasets

4.1.1 Financial Phrase Bank

Financial Phrase Bank (FPB) (Malo et al., 2014) is one of the most frequently used and cited datasets in the financial sentiment analysis field. It has also been used to fine-tune the financial language models used for attack in this thesis. It consists of financial news on all listed companies in OMX Helsinki downloaded from the LexisNexis database using an automated web scraper. They are sampled randomly from a subset of 10000 articles to obtain good coverage across small and large companies in various industries from multiple sources. Following filtering out the sentences which do not include lexicon entities and random sampling, 5000 sentences have been selected for annotation. To ensure high-quality annotations, each text sample is categorised into the three sentiment classes (i.e., negative, neural and positive) that we use in this thesis by 5-8 annotators with backgrounds in finance and business. As there are multiple annotators, the dataset comes in multiple sizes depending on the percentage of agreeing annotators: 50agree, 66agree, 75agree, and 100agree. We used the 100agree subset this is expected to lead to more decisive performance results for our adversarial attack approach. This subset includes 2264 sentences, consisting of 570 positive, 303 negative, and 1391 neutral sentences. Fig. 4.1 presents the distribution (histogram) of the samples over the sentiment classes for the FPB and the two remaining datasets described in the subsequent sections. The neutral class emerges as the dominant category, likely due to the nature of financial news and tweets, which are predominantly neutral under stable economic conditions, apart from brief periods of economic booms and busts.

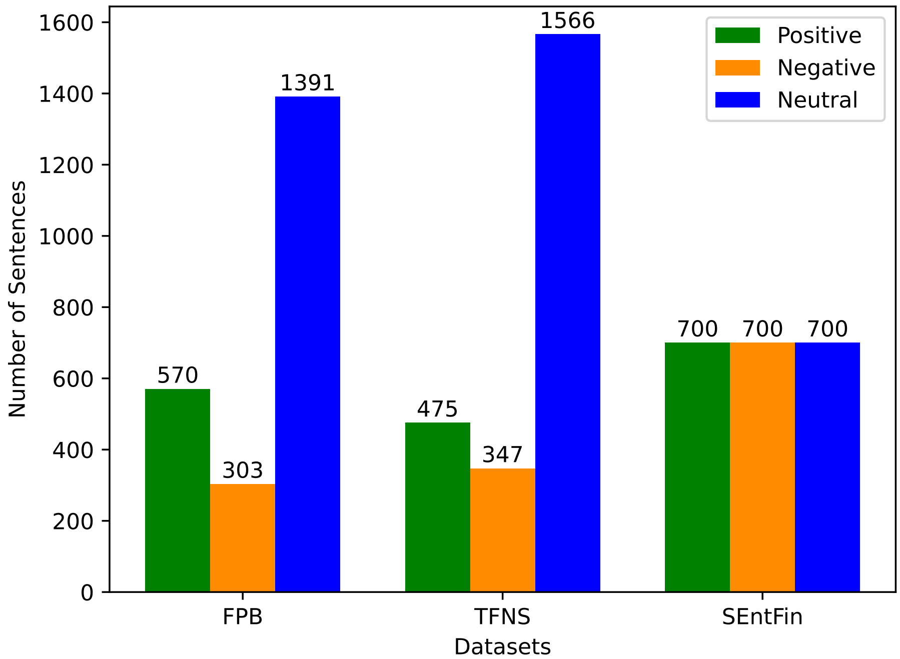


Figure 4.1 – Distribution of samples by sentiment class for the three datasets.

4.1.2 Twitter Financial News Sentiment

Twitter Financial News Sentiment dataset (TFNS) (Goin, 2022) comprises tweets sourced from a various origins, including media outlets and individual investors, all containing financial terminology such as stock tickers, indices, and common financial abbreviations. The dataset encompasses 9540 text samples designated for training and 2388 samples for validation. We used the validation split because none of the attacked models are pre-trained or fine-tuned on it, and it has a comparable sample size to the FPB dataset. Each sample is annotated with one of three investment-related sentiments: bearish, bullish, or neutral. For the purposes of our analysis, we categorized bearish sentiment as negative and bullish sentiment as positive. Similar to the FPB dataset, a significant proportion of the sentences are labeled as neutral for the same reason mentioned in the previous section.

4.1.3 Sentiment and Entity Annotated Financial News Dataset

Sentiment and Entity annotated Financial News dataset (SEntFiN) (Sinha et al., 2022) comprises 10753 news headlines covering 1009 entities across multiple sectors and companies, sourced from The Economics Times, an Indian business newspaper. This human-annotated dataset is among the largest available and addresses the entity-based sentiment analysis problem. This analysis entails identifying and extracting entities within headlines, followed by classifying the sentiment associated with each entity. Consistent with the other two datasets, sentiments are classified into three categories: positive, negative, and neutral. The annotation process involves three annotators working collaboratively to classify each sample, adopting a consensus-driven approach to enhance reliability. In instances where consensus is not reached, a curator intervenes to make the final determination. Notably, over 98% of the dataset entries feature labels with unanimous annotator agreement.

We preprocessed the SEntFiN dataset to ensure compatibility with our adversarial attack methodology, which requires a single sentiment label per sample. However, due to the annotation process, a text sample involving multiple entities might exhibit conflicting sentiment labels. To address this, we discarded samples with multiple non-uniform sentiment labels, resulting in a refined set of 9514 samples. From this remaining set, we created a class-balanced dataset comprising 2100 samples using random sampling without replacement. That is why, Fig. 4.1 exhibits a flat distribution over the sentiment classes for this dataset.

4.2 Language Models

In this thesis, we take two leading financial sentiment classification models that have been extensively evaluated across various financial benchmarks. Our objective is to assess the robustness of their performance against adversarial attacks and identify their vulnerabilities. To achieve this, we used ChatGPT, an advanced general-purpose large language model, to generate the sample paraphrases and perform the sentiment verifications presented in Chapter 3.

In the following, we first describe the two financial models that were attacked and then the advanced one used for creating these attacks.

4.2.1 FinBERT

FinBERT (Araci, 2019) is a financial sentiment analysis model based on BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018). It is further pre-trained on an unlabeled financial domain-specific corpus called TRC-2, which contains approximately 29 million words and around 400,000 sentences. Following this pre-training, the model is fine-tuned on 80% of the 50agree subset of the FBP dataset for the sentiment classification task.

FinBERT is primarily designed as a single-task model, which remains one of the state-of-the-art models, demonstrating high performance on benchmark datasets such as FPB, as well as Financial Opinion Mining and Question Answering (FiQA) (Maia et al., 2018). Notably, it outperforms BloombergGPT (Wu et al., 2023), a 50-billion parameter multi-task language model tailored to the financial domain, on the sentiment classification task.

We obtained the embeddings for this model from the output of the pooling layer (i.e., pooled_output). This layer consists of a linear layer followed by tanh activation and takes the CLS token representation from the FinBERT model as input. The pooling output is then passed through another linear layer followed by softmax activation to produce the final classification probabilities for the three sentiment classes. We normalize the 768-dimensional embeddings to compute cosine similarity over the resulting unit-vector representations.

4.2.2 FinGPT

FinGPT (Yang, Liu, et al., 2023) is a framework targeting to provide open-source financial LLMs using various backbone pre-trained models. We used the FinGPT v3.2 version, which is based on the LLaMa2 7-billion model (Touvron et al., 2023). According to the authors, this version is one of the highest-performing models in benchmarks (AI4Finance-Foundation, 2023a).

The model is fine-tuned simultaneously on multiple financial datasets and tasks comprising in financial sentiment analysis, relation extraction, headline classification, and named entity recognition. More specifically, for the sentiment classification task, the FinGPT model is trained on (AI4Finance-Foundation, 2023b) a random split of 75% of the FPB dataset (the 50agree set) (Malo et al., 2014), all the training samples of the TFNS dataset (Goin, 2022), 22.6% of the FiQA dataset, and finally, the train set of the News with GPT instructions (NGI) dataset (Wang, 2023).

In contrast to the BERT-based models, which employ the CLS token and a pooling layer to extract meaningful sentence embeddings, LLaMa-based ones lack a straightforward way of extracting such embeddings to be used in downstream classification tasks. To address this limitation, we tried the AnglE method (Li and Li, 2023), which minimizes a multi-loss objective function to avoid the common problem of vanishing gradients in extracting text embeddings. For LLaMa-based models, this is done by fine-tuning a base model with the specific prompt "Summarize sentence {sentence} in one word:" while using the multi-loss objective. Although being effective for stand-alone text embedding generation tasks, this approach gave sub-optimal results in our experiments due to misalignment of the FinGPT and the AnglE-optimized models, which differ in terms of the both datasets and loss functions they use in fine-tuning the LLaMa2 base model.

To address this issue, we conducted experiments on pairs of both similar (paraphrased) and unrelated sentences. We observed that, although FinGPT effectively captures semantic information, its embedding similarity values consistently fall within a narrow range of [0.95, 1.0] across all tested pairs. Motivated by this finding, we devised a simple strategy to calibrate these similarity values on a dataset. It involves learning a linear mapping to transform (extend) them to the range [0.0, 1.0].

More specifically, we first randomly select a fixed number (2000) of sample pairs from the FPB dataset and compute a lower bound l_{simil} for the embedding similarity over all of them. To extract the embeddings from the FinGPT model, we use the last hidden state and normalize it to have unit magnitude vectors. We then learn the linear mapping that transforms this lower bound l_{simil} to 0.0, while keeping the upper bound at 1.0. This calibration step produces a scaling factor and a bias value, which we apply to the embedding similarity values - clipping them to ensure they always remain within the range $[0, 1]$ - during the generation of adversarial samples described in Chapter 3. The resulting l_{simil} value obtained was 0.967, which is used in all our experiments involving the FinGPT model.

4.2.3 ChatGPT

ChatGPT, developed by OpenAI, is a family of generative large language models that has attracted considerable attention for its robust performance across a wide range of NLP tasks, including text generation, conversation, and language understanding. Recent GPT versions demonstrate the capability to produce contextually appropriate and nuanced text in complex and intricate situations (Achiam et al., 2023), a feature essential for generating adversarial samples within a financial context.

In this thesis, we employed GPT-4o (omni) (OpenAI, 2024b), the latest model from the family of OpenAI's LLMs, to generate adversarial attacks targeting FinBERT and FinGPT. After conducting experimental trials with GPT-3.5, GPT-4 and GPT-4o, we selected GPT-4o for its superior semantic quality in rephrasing complex financial text compared to GPT-3.5, and its cost-effectiveness relative to GPT-4. GPT-4o's capability to produce human-like text is of particular significance as it is needed to generate natural adversarial samples.

To evaluate the embedding similarity between the paraphrased sentences and the original ones, we utilized the most recent embedding model, called 'text-embedding-3-large' (OpenAI, 2024a). This approach was necessitated by the inability to obtain embeddings directly from the GPT4-o model as of 2024. However, as demonstrated in the following chapter, this embedding model was sufficiently effective in generating strong adversarial attacks on both financial models.

Chapter 5

Results

In this chapter, we present the results of our adversarial attack approach on the FinBERT and FinGPT models. The primary performance metrics used in this analysis include flip rate and mean absolute error. Given a dataset, the flip rate is defined as the proportion of samples for which the attacked model changes its original prediction on a text sample after being presented with a perturbed (paraphrased) version of that sample.

Let $\hat{y}_{\text{orig}}(i)$ and $\hat{y}_{\text{attack}}(i)$ denote the sentiment predictions of the attacked model for the i-th original sample from the dataset and its generated adversarial sample, respectively. The flip rate (FR) is defined as:

$$\text{FR} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}(\hat{y}_{\text{orig}}(i) \neq \hat{y}_{\text{attack}}(i)) \tag{5.1}$$

where M is the total number of samples in the dataset, and $\mathbf{1}(\cdot)$ is the indicator function that returns 1 if the given condition is true, and 0 otherwise. *Mean Absolute Error (MAE)* is defined in a similar way but instead of using an indicator function which assigns an equal (unity) weight to each flip (or transition), MAE assigns a higher weight for transitions between the positive and negative classes. It is defined as:

$$\text{MAE} = \frac{1}{M} \sum_{i=1}^{M} |\hat{y}_{\text{orig}}(i) - \hat{y}_{\text{attack}}(i)| \tag{5.2}$$

The predictions $\hat{y}_{\text{orig}}(i)$ and $\hat{y}_{\text{attack}}(i)$ can take values from the set $\{-1, 0, 1\}$, where -1, 0, and 1 respectively correspond to negative, neutral, and positive classes. MAE imposes a greater penalty for sentiment transitions between the positive and negative classes compared to the two remaining class pairs (Fatouros et al., 2023).

We employ FR and MAE as primary metrics since ground truth labels are unavailable in many realistic adversarial attack scenarios. We aim to ascertain the extent to which we can confound the model's assessment of financial sentiment through our adversarial attack, which is based solely on embedding-similarity optimization.

We start our analysis by evaluating the performance of GPT-4o, the model used to generate adversarial perturbations in this thesis, across the three datasets employed. We applied a zero-shot methodology and conducted three separate trials on each dataset to mitigate the influence of randomness in the outputs produced by GPT-4o. In each trial, we queried the sentiment of each sentence three times, setting both the temperature and top_p parameters to 0.1 to minimize variability; this is crucial as the task of financial sentiment classification typically does not involve inherent randomness. The final sentiment label for each sentence was determined by the majority label across the trials.

Table 5.1 summarizes the results over the three sentiment classification datasets. Despite being in zero-shot setting, GPT-4o demonstrates notably high performance in the classification task. As will be evidenced in the following section, it achieves results comparable to FinBERT and FinGPT on the FPB dataset, despite both models being fine-tuned on this dataset. However, on the remaining two datasets, which were unseen by both models, GPT-4o significantly outperforms them. This underscores the suitability of GPT-4o for generating adversarial samples involving nuanced financial language and terminology.

Table 5.1 – GPT-4o zero-shot performance on the three datasets.

	FPB	**TFNS**	**SEntFiN**
Accuracy	0.96	0.80	0.85
F1 Score	0.96	0.77	0.85
Precision	0.96	0.74	0.85
Recall	0.96	0.84	0.85

5.1 Performance Results

In this section, we present the results of embedding-similarity optimization-based adversarial attacks on FinBERT and FinGPT over the three datasets.

The results shown in Table 5.2 demonstrate that the proposed method substantially impacts the sentiment predictions generated by both models. Despite being fine-tuned on the FPB dataset, they exhibit flip rates of 30% and 22% respectively, highlighting the effectiveness of the proposed attack methodology.

Table 5.2 – Results of the proposed adversarial attack approach on the FinBERT and FinGPT models over the three datasets used in this thesis.

	FinBERT						**FinGPT**					
	FPB		**TFNS**		**SEntFiN**		**FPB**		**TFNS**		**SEntFiN**	
Metrics	Before Attack	After Attack	Before Attack	After Attack	Before Attack	After Attack	Before Attack	After Attack	Before Attack	After Attack	Before Attack	After Attack
FR	-	0.304	-	0.540	-	0.505	-	0.217	-	0.426	-	0.399
MAE	-	0.388	-	0.657	-	0.669	-	0.236	-	0.483	-	0.467
Accuracy	0.972	0.710	0.725	0.531	0.721	0.580	0.992	0.785	0.761	0.620	0.791	0.680
F1 Score	0.963	0.669	0.668	0.482	0.723	0.573	0.990	0.763	0.751	0.592	0.786	0.681
Precision	0.951	0.659	0.652	0.487	0.732	0.584	0.990	0.749	0.723	0.578	0.795	0.683
Recall	0.976	0.708	0.705	0.540	0.721	0.580	0.989	0.781	0.846	0.662	0.791	0.680

The data presented in Table 5.2 reveal a significant deviation between the FR and the MAE metric values for the FinBERT model. This deviation is further evidenced by Table 5.3, which shows that the percentage of transitions between the positive and negative classes remain above 20% across the three datasets. For a given pair of sentiment classes (i, j), this transition percentage T_{ij} is calculated as

$$T_{ij} = \left(\frac{M_{ij} + M_{ji}}{K_i + K_j} \right) \times 100 \tag{5.3}$$

where M_{ij} is the number of predictions transitioning from class i to j, and K_i is the total number of samples initially predicted in sentiment class i before the attack.

Analysis of Table 5.3 indicates that the bidirectional transitions between the neutral and positive classes are the highest among all datasets for both models, except for FinBERT on FPB, for which the transition percentage of the positive-negative pair is slightly higher. Notably, the dominance

Table 5.3 – Percentage of transitions between pair of sentiment classes pre-and post-attack.

	FinBERT			FinGPT		
Sentiment Pairs	**FPB**	**TFNS**	**SEntFiN**	**FPB**	**TFNS**	**SEntFiN**
Positive ↔ Negative	20.7%	27.9%	27.4%	4.9%	10.5%	9.2%
Negative ↔ Neutral	6.1%	22.6%	16.7%	6.3%	18.6%	17.3%
Neutral ↔ Positive	20.5%	31.1%	32.2%	17.4%	31.0%	34.1%

of the neutral-positive transitions is also visible from the inter-annotator disagreements of the FPB dataset, for which the leading pair for disagreement is the same (Malo et al., 2014). For the FBP dataset, this is primarily due to the significant number of samples exhibiting borderline sentiment, falling between positive and neutral.

To gain additional insights, we adapt the concept a confusion matrix to define a *flip confusion matrix*. This matrix serves as similar analytical role but differs in that it compares the model's predictions before and after the attack, rather than comparing the ground truth labels to the post-attack predictions. This analysis allows us to observe the directional transition in the predictions of FinBERT and FinGPT before and after the attack.

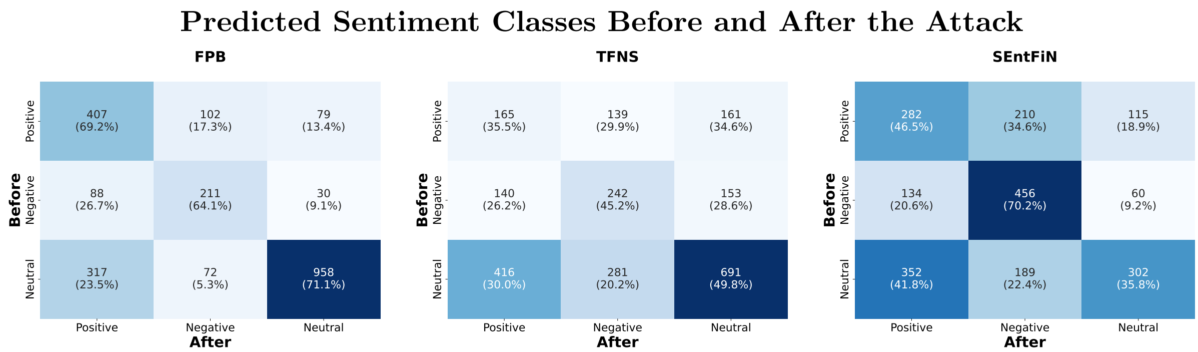


Figure 5.1 – Flip confusion matrices depicting the shift in FinBERT's sentiment predictions for three datasets before and after exposure to an adversarial attack. Each row in the matrices tracks the transition of predictions from a specific initial class (positive, negative, neutral) to the resultant classes following the adversarial attack.

Fig. 5.1 presents the flip confusion matrices for the FinBERT model for the three datasets. The percentage value in parenthesis within each cell is calculated as the proportion of the single-direction sentiment transitions for each pre-attack sentiment class (row). They can therefore be viewed as normalized transition rates. We observe that transitions from neutral to positive are more common than those from positive to neutral in both the FPB and SEntFiN datasets. In contrast, TFNS

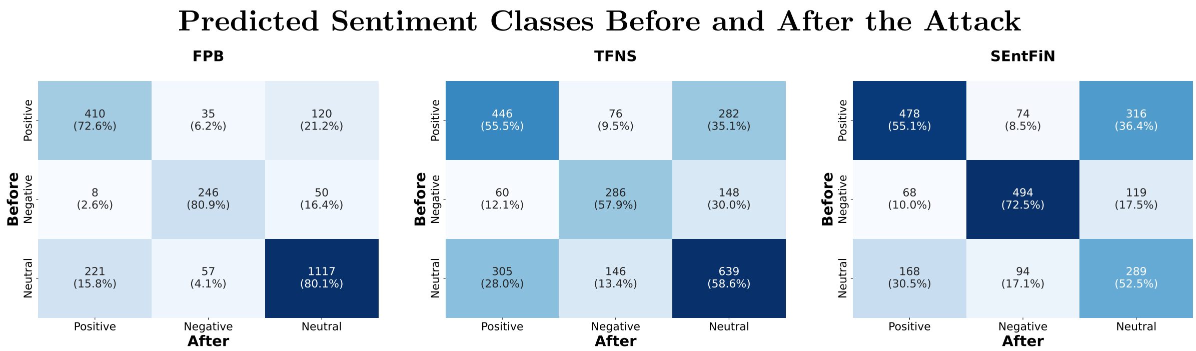


Figure 5.2 – Confusion matrices depicting the shift in FinGPT's sentiment predictions for three datasets before and after exposure to an adversarial attack. Each row in the matrices tracks the transition of predictions from a specific initial class (positive, negative, neutral) to the resultant classes following the adversarial attack.

dataset has a slightly higher percentage of transitions from positive to neutral.

Fig. 5.2 depicts the flip confusion matrices for FinGPT. They reveal a prominent direction across all there datasets, which is from positive to neutral. This strikingly contrasts FinBERT's dominant transition direction.

In Table 5.2, in addition to presenting the flip rate and MAE as primary results, we also provide for both models the accuracy, macro F1-score, precision and recall metrics using the ground truth labels present in the datasets used. Due to the larger size of its baseline model and the wide extent of the datasets it was trained on, the FinGPT model outperforms FinBERT on the financial sentiment analysis task. However, despite our main goal is to alter the classifiers' output predictions, our adversarial attack approach reduces the accuracy of both models for more than 20% on the FBP dataset they were fine-tuned on. More specifically, the accuracy of the FinBERT and FinGPT models drop respectively from 97.2% and 99.2% to 71.0% and 78.5% on this dataset. On TFNS and SEntFin, due to the significantly lower base accuracy of the two models (i.e., lower and upper seventies for FinBERT and FinGPT respectively), the drop in accuracy remain in the range $[10\%, 20\%]$. Additional results comprising confusion matrices for FinGPT and FinBERT, are provided in Appendix A.1 and Appendix A.2, respectively.

5.2 Vulnerabilities

In this section, we briefly explain the identified vulnerabilities in the FinBERT and FinGPT models examining original samples of the FPB dataset and the adversarial ones we generated using our approach. The results of this analysis are presented in Table 5.4 for FinBERT and in Table 5.5 for FinGPT.

The cases where models failed to maintain their sentiment against adversarial samples are classified into three categories, one of which involves difficulties in numeric comprehension. As mentioned in the work of FinBERT (Araci, 2019), this model struggles to compare numbers and assess their impact on financial sentiment in the absence of explicit directional words such as 'increase,' 'rise,' 'up,' 'down,' 'decrease,' etc. This limitation is demonstrated through its predictions on adversarial samples containing numbers without directional statements that carry sentiment-related information. On a large majority of these adversarial samples, FinBERT fails to maintain its initial prediction. Similarly, FinGPT exhibits the same limitation; its predictions can be easily altered on adversarial samples, where the sentiment is solely carried through numerical comparison. For instance, when the directional statement 'down from' is removed from the original sample (1a) in Table 5.5, the model's prediction switches from negative to neutral, despite the sentiment of the adversarial sample (1b) remaining the same.

Another revealed weakness in both models pertains to their handling of key verbs that carry negative sentiments within double negative statements. These models often misinterpret the sentiment when presented with such adversarial samples. Typically, verbs like 'decrease', 'drop', and 'fall' are associated with negative sentiments. However, when they are used in contexts such as 'financial losses', 'costs', or 'job cuts', the implied sentiment should be positive due to the beneficial financial implications of the reductions. Contrary to this expectation, when FinBERT and FinGPT encounter adversarial samples that include these terms, particularly in contexts with strong negative connotations, they erroneously switch their sentiment to negative.

Another major vulnerability observed for both models is their sensitivity to specific words, which can significantly impact sentiment predictions. The presence of these words often results in predictions that disregard the broader context in which they are used. For instance, when the phrase 'fall

Vulnerability Category	**Original Input (a) and Adversarial (b) Samples**	**Pre-And Post-Attack Predictions**
Numerical Comprehension	1a. Operating profit was EUR 11.4 mn, **up from** EUR 7.5 mn .	positive
	1b. Previously at EUR 7.5 million, operating profit reached EUR 11.4 million.	neutral
	2a. In August-October 2010 , the company 's result before taxes totalled EUR 9.6 mn , **up from** EUR 0.5 mn in the corresponding period in 2009 .	positive
	2b. The pre-tax profit for the company in August-October 2010 was EUR 9.6 million, compared to EUR 0.5 million during the same timeframe in 2009.	negative
	3a. Finnish electronics contract manufacturer Scanfil had net sales of EUR 52.2 mn in the first quarter of 2007 , **down from** EUR 60.1 mn a year before .	negative
	3b. During the first quarter of 2007, Finnish firm Scanfil had net sales amounting to EUR 52.2 million, compared to EUR 60.1 million one year before.	positive
Use of Key Verbs (e.g., drop,fall) in Double Negative Statements	4a.Metsaliitto , however , narrowed its net loss for the second quarter of 2007 to 5.0 mln euro $ 6.9 mln from 61 mln euro $ 83.7 mln a year ago .	positive
	4b. Metsaliitto's **net loss fell to** €5.0 million ($6.9 million) in the second quarter of 2007, from €61 million ($83.7 million) a year earlier over 2007.	negative
	5a.For the first nine months of 2010 , Talvivaara 's net loss narrowed to EUR 8.3 million from EUR 21.9 million for the same period of 2009.	positive
	5b. Talvivaara's **net loss** for the period from January to September 2010 **fell to** EUR 8.3 million, compared to EUR 21.9 million for the same period in 2009.	negative
	6a. Finnish metal products company Componenta Oyj (HEL : CTH1V) said today its net loss narrowed to EUR 500,000 (USD 680,000) in the last quarter of 2010 from EUR 5.3 million for the same period a year earlier .	positive
	6b. The **net loss** of Componenta Oyj (HEL: CTH1V), a Finnish metal products firm, **dropped to** EUR 500,000 (USD 680,000) in the last quarter of 2010 compared to EUR 5.3 million in the previous year's similar quarter.	negative
	7a. Finnish steel maker Rautaruukki Oyj (HEL : RTRKS) , or Ruukki , said today its net loss contracted to EUR 49 million (USD 68.2 m) for the first nine months of 2010 from EUR 229 million for the same period a year ago .	positive
	7b. Ruukki, officially known as Rautaruukki Oyj, reported today its **net loss** for January through September 2010 **dropped to** EUR 49 million (USD 68.2 million), compared with EUR 229 million the year earlier.	negative
Sensitivity to Specific Words	8a. BAVARIA Industriekapital AG 's 2006 revenues were EUR 333 million , with an EBITDA of EUR 51 million .	neutral
	8b. With EUR 333 million in revenues, BAVARIA Industriekapital AG's 2006 EBITDA **reached** EUR 51 million.	positive
	9a. The company expects its net sales for the whole of 2007 to be EUR 950mn-1 ,000 mn.	neutral
	9b. For the year 2007, the company anticipates net sales revenue **to fall between** EUR 950 million and EUR 1 billion.	negative
	10a. The total value of these two contracts is over EUR 21 million .	neutral
	10b. The overall worth of these two contracts is **higher than** EUR 21 million.	positive

Table 5.4 – Examples of identified vulnerabilities in the FinBERT model. Original input samples (a) are from the FPB dataset, while adversarial samples (b) are generated by our approach. The model's predictions on the original inputs match the ground truth labels.

within the range' is used to express a numerical range, FinBERT mistakenly shifts its predictions to negative. Similarly, FinGPT exhibits a comparable sensitivity; despite the context clearly expressing a positive outcome, the inclusion of the word 'hit' in the adversarial samples often causes its sentiment to switch to negative, as demonstrated in sample (6a-b) given in Table 5.5.

Vulnerability Category	**Original Input (a) and Adversarial (b) Samples**	**Pre-And Post-Attack Predictions**
Numerical Comprehension	1a. Sampo Bank 's market share of lending was 13.6 % , **down from** 14.4 % in the first quarter of 2008 .	negative
	As compared to 14.4% in Q1 2008, Sampo Bank's current lending market share is 13.6%.	neutral
	2a. Net interest income was EUR 152.2 mn , **up from** EUR 101.0 mn in 2008 .	positive
	2b. In comparison to the EUR 101.0 million in 2008, net interest income reached EUR 152.2 million.	neutral
Use of Key Verbs (e.g., decrease,down) in Double Negative Statements	3a.Finnish fibers and plastic products maker Suominen Corporation said its net loss narrowed to 1.6 mln euro $ 2.0 mln in the first nine months of 2006 from 2.16 mln euro $ 2.7 mln in the same period of 2005 .	positive
	3b. Suominen Corporation of Finland recorded a **net loss decrease** to 1.6 million euros (2.0 million USD) in January through September of 2006 from 2.16 million euros (2.7 million USD) in the identical period of the previous year.	negative
	4a.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 .	positive
	4b. Raute's **net loss** in the second quarter of 2010 was EUR 123,000, **down from** EUR 1.5 million in the corresponding period in 2009.	negative
Sensitivity to specific words	5a. The subsidiary will be responsible for filter sales , local assembly of filters and after market services in China .	neutral
	5b. Sales of filters, local filter assembly, and after-sale services in China will **fall under** the subsidiary's purview.	negative
	6a. Pre-tax profit totaled EUR 397.4 mn , up from EUR 164.7 mn .	positive
	6b. Pre-tax profit **hit** EUR 397.4 million, climbing from EUR 164.7 million.	negative
	7a. The acquisition was financed with $ 2.56 billion of debt arranged by Goldman , Sachs & Co. .	neutral
	7b. Goldman, Sachs & Co. **secured** $2.56 billion in debt to support the acquisition.	positive

Table 5.5 – Examples of identified vulnerabilities in the FinGPT model. Original input samples (a) are from the FPB dataset, while adversarial samples (b) are generated by our approach. The model's predictions on the original inputs match the ground truth labels.

Although both models exhibit these vulnerabilities, in our analysis, we observed that FinBERT is more prone to them, particularly in terms of sensitivity to key verbs in double negative statements and specific words in the samples. This heightened sensitivity often leads to more transitions between positive and negative sentiments for FinBERT, as shown in Table 5.3 of Section 5.1. In contrast, FinGPT exhibits significantly fewer failure cases, and considerably less frequent transitions between positive and negative sentiments. These findings account for the lower overall MAE scores presented in Table 5.2 for FinGPT compared to FinBERT.

Chapter 6

Conclusion

In this thesis, we have investigated the vulnerability of financial sentiment analysis models to adversarial attacks and introduced a novel, white-box attack methodology leveraging a pre-trained general-purpose language model (GPT-4o). Our approach is distinct in that it does not require training a task-or domain-specific adversarial model or utilizing computationally intensive gradient-based optimization techniques. Instead, it generates high-quality, semantically consistent adversarial samples by employing carefully crafted instructions and a simple embedding-similarity-driven optimization algorithm.

Our experimental results demonstrate that both FinBERT and FinGPT, leading models in financial sentiment analysis, exhibit significant susceptibility to our proposed adversarial attacks. Specifically, the sentiment predictions of these models were successfully altered for a substantial proportion of the samples across three public datasets, including Financial Phrase Bank (FPB), Twitter Financial News Sentiment (TFNS), and Sentiment and Entity Annotated Financial News (SEntFiN).

The flip rate (FR), a metric we defined to measure the proportion of samples for which the model's prediction changes after an adversarial attack, was 30% for FinBERT and 22% for FinGPT on the FPB dataset, highlighting the efficacy of our approach even over a dataset they were fine tuned on. Moreover, our approach achieves a mean absolute error (MAE) of more than 0.65 for FinBERT and 0.45 for FinGPT on the remaining two datasets, respectively. This highlights the significant impact of our attacks on the predictions of these models, with the most prominent sentiment transitions occurring between neutral and positive classes.

Our findings emphasize the need for enhanced robustness in financial classification models against adversarial attacks. By understanding and addressing these vulnerabilities, it is possible to improve the reliability and security of automated financial systems. Future research could explore integrating adversarial training techniques to bolster model resilience and develop more sophisticated attack methodologies that encompass a wider range of perturbations.

Future work could also involve applying the proposed approach to multiple classification models simultaneously, predicting an attack that can deceive a group of models rather than a single one. Additionally, employing multiple advanced models, including both generalist (language) and specialist (finance) ones, could create a richer set of attacks. We believe the proposed approach is easily extendable to both of these scenarios.

Appendices

Appendix A

Confusion Matrices

A.1 Confusion Matrices for FinBERT

Figure A.1 – Confusion matrices obtained for the FinBERT model on the FPB dataset.

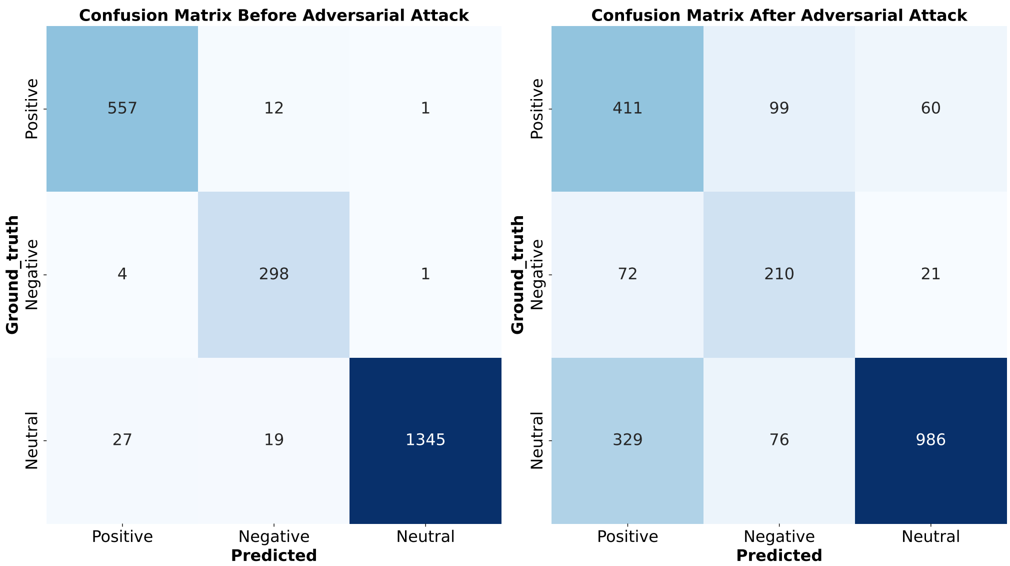


Figure A.2 – Confusion matrices obtained for the FinBERT model on the TFNS dataset.

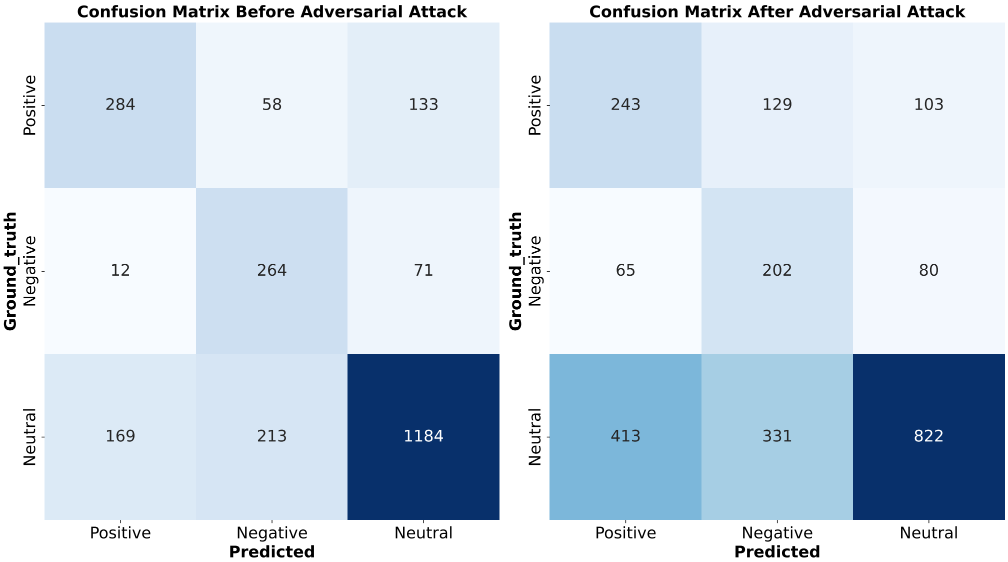
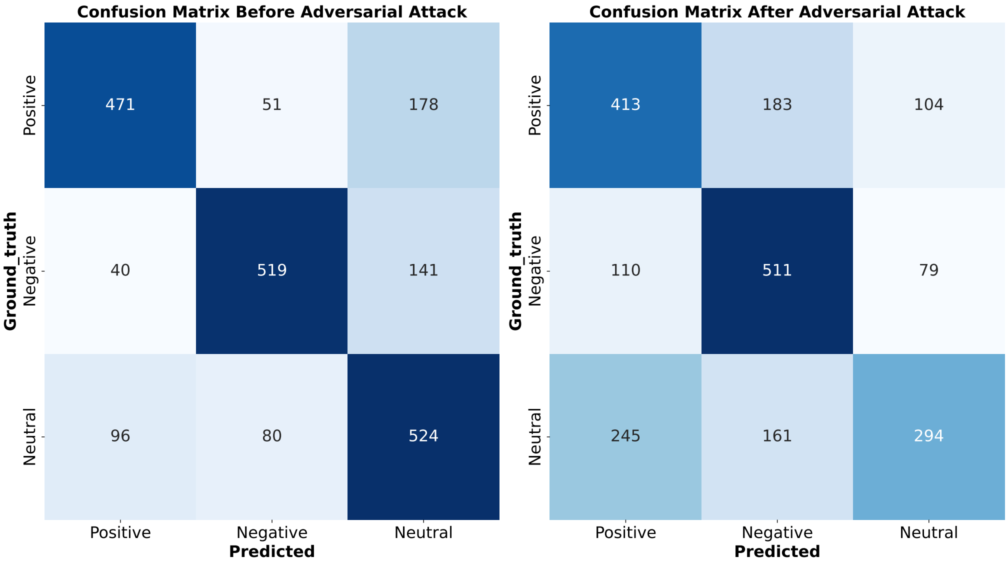


Figure A.3 – Confusion matrices obtained for the FinBERT model on the SEntFiN dataset.



A.2 Confusion Matrices for FinGPT

Figure A.4 – Confusion matrices obtained for the FinGPT model on the FPB dataset.

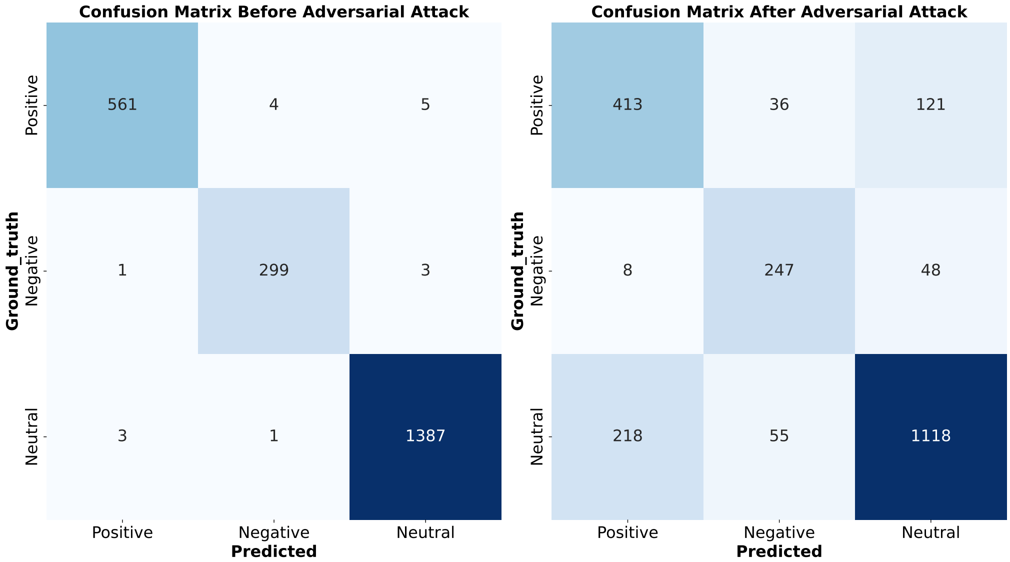


Figure A.5 – Confusion matrices obtained for the FinGPT model on the TFNS dataset.

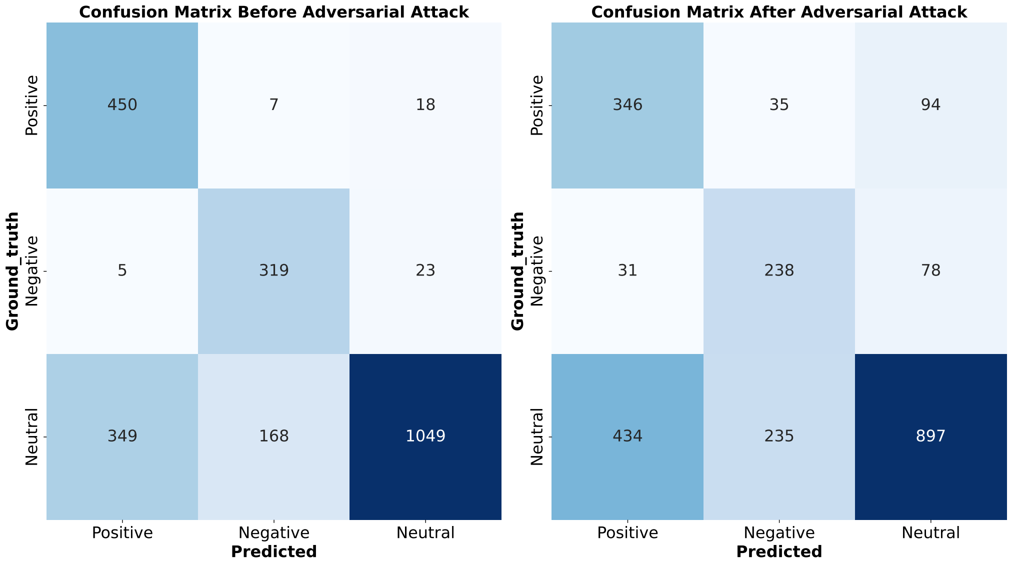
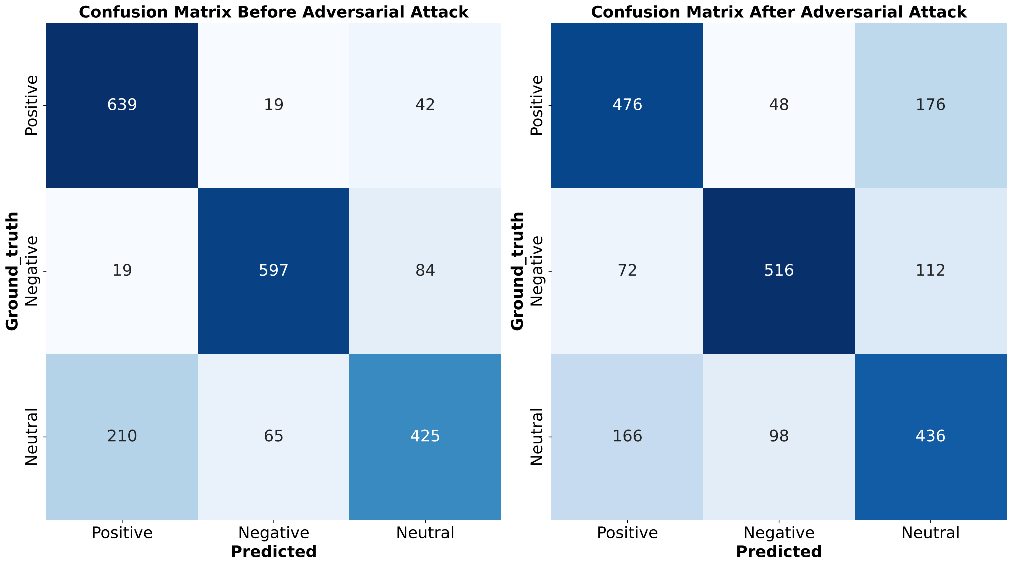


Figure A.6 – Confusion matrices obtained for the FinGPT model on the SEntFiN dataset.



Appendix B

Examples of Generated Paraphrases

Table B.1 – Examples of generated paraphrases for an input sample with positive ground-truth label. Sample 0 is the original text from FPB, and samples 1-25 are the GPT-4o generated paraphrases.

Number	**Original Sentence (0) and Generated Paraphrases (1-25)**
0	Sales rose 10 pct to 566 mln eur on the back of strong volume and favourable currency effects.
1	Sales increased by 10 percent to 566 million euros due to high volume and positive currency impacts.
2	A 10% rise in sales brought them to 566 million euros, driven by strong volume and advantageous currency shifts.
3	The sales figure surged by 10% to 566 million euros, fueled by robust volume and beneficial currency movements.
4	Sales saw a 10% uplift to 566 million euros, supported by strong volume and favorable currency fluctuations.
5	There was a 10 percent increase in sales, reaching 566 million euros, owing to strong volume and positive currency effects.
6	Thanks to high volume and favorable currency effects, sales climbed 10% to 566 million euros.
7	Sales boosted by 10 percent to 566 million euros on account of strong volume and advantageous currency effects.
8	The 10 percent rise in sales took them to 566 million euros, aided by high volume and positive currency effects.
9	With strong volume and favorable currency effects, sales rose 10% to reach 566 million euros.
10	Sales jumped 10% to 566 million euros due to strong volume and advantageous currency movements.
11	A 10% increase in sales resulted in a total of 566 million euros, driven by strong volume and favorable currency changes.
12	Sales advanced by 10 percent to 566 million euros, helped by robust volume and beneficial currency shifts.
13	Thanks to favorable currency movements and strong volume, sales rose 10% to 566 million euros.
14	Sales grew 10% to 566 million euros, underpinned by high volume and positive currency movements.
15	The 10 percent increase in sales brought the total to 566 million euros, boosted by strong volume and favorable currency impacts.
16	A rise of 10% in sales to 566 million euros was seen, driven by strong volume and advantageous currency fluctuations.
17	Fueled by strong volume and favorable currency, sales rose 10% to 566 million euros.
18	The sales figure jumped by 10 percent to 566 million euros, thanks to strong volume and advantageous currency effects.
19	Sales ascended 10% to 566 million euros due to strong volume and favorable currency dynamics.
20	On the back of strong volume and favorable currency effects, sales climbed 10% to 566 million euros.
21	Sales improved by 10% to 566 million euros, due to robust volume and positive currency effects.
22	A 10% boost in sales raised the total to 566 million euros, driven by high volume and favorable currency shifts.
23	Thanks to both robust volume and positive currency impact, sales increased 10% to 566 million euros.
24	Sales surged 10% to 566 million euros, owing to strong volume and advantageous currency environment.
25	Driven by strong volume and favorable currency effects, sales went up 10% to 566 million euros.

Table B.2 – Examples of generated paraphrases for an input sample with neutral ground-truth label. Sample 0 is the original text from FPB, and samples 1-25 are the GPT-4o generated paraphrases.

Number	**Original Sentence (0) and Generated Paraphrases (1-25)**
0	Tiimari operates 194 stores in six countries – including its core Finnish market – and generated a turnover of 76.5 mln eur in 2005.
1	Tiimari runs 194 outlets across six nations, including its primary market of Finland, achieving 76.5 million euros in turnover in 2005.
2	In 2005, Tiimari operated 194 retail locations across six countries, with Finland being the main market, generating 76.5 million euros in sales.
3	With 194 stores in six countries, Tiimari's core market is Finland, and it had a turnover of 76.5 million euros in 2005.
4	Tiimari boasted 194 stores in 2005, spanning six countries, and its primary market, Finland, contributed to a 76.5 million euro turnover.
5	Operating 194 stores in six nations, Tiimari focused on Finland in 2005, resulting in a 76.5 million euro turnover.
6	In six countries, including Finland, Tiimari's 194 stores achieved a turnover of 76.5 million euros in 2005.
7	Tiimari's operations in 194 stores across six countries, with Finland as the core market, produced a turnover of 76.5 million euros in 2005.
8	The Finnish market was central for Tiimari's 194 stores in 2005, operating in six countries and generating 76.5 million euros in turnover.
9	Tiimari operated 194 outlets in six countries in 2005, focusing on Finland, and garnered a 76.5 million euro turnover.
10	In 2005, Tiimari's 194 stores across six countries, including its main Finnish market, totaled a turnover of 76.5 million euros.
11	With Finland as its core, Tiimari managed 194 stores in six countries in 2005, achieving a 76.5 million euro turnover.
12	Tiimari's 194 stores across six countries, with Finland being its core market, earned a turnover of 76.5 million euros in 2005.
13	Six countries, including Finland, housed Tiimari's 194 stores, which generated 76.5 million euros in turnover in 2005.
14	Tiimari achieved a 76.5 million euro turnover in 2005 with its 194 stores across six countries, centering on Finland.
15	In 2005, Tiimari's 194 stores, spread over six countries including its main Finnish market, posted a turnover of 76.5 million euros.
16	With Finland as the core, Tiimari operated 194 stores across six nations in 2005, achieving a turnover of 76.5 million euros.
17	Tiimari's 194 retail locations in six countries, with Finland as the main market, resulted in a 76.5 million euro turnover in 2005.
18	Across six countries, Tiimari's 194 stores, including the key Finnish market, reported a turnover of 76.5 million euros in 2005.
19	In 2005, Tiimari's 194 stores across six nations generated a turnover of 76.5 million euros, with Finland as the main focus.
20	Tiimari's operations in six countries, with 194 stores and a key focus on Finland, recorded a 76.5 million euro turnover in 2005.
21	Tiimari had 194 stores in six countries, including its crucial Finnish market, and achieved a turnover of 76.5 million euros in 2005.
22	In 2005, Tiimari's core Finnish market and other five countries housed 194 stores, yielding a turnover of 76.5 million euros.
23	Tiimari, operating 194 stores in six countries, focused on Finland in 2005, recording a turnover of 76.5 million euros.
24	The core Finnish market and five other countries saw Tiimari operate 194 stores, with a 2005 turnover of 76.5 million euros.
25	Tiimari ran 194 stores in six nations in 2005, including Finland at its core, reaching a turnover of 76.5 million euros.

Table B.3 – Examples of generated paraphrases for an input sample with negative ground-truth label. Sample 0 is the original text from FPB, and samples 1-25 are the GPT-4o generated paraphrases.

Number	Original Sentence (0) and Generated Paraphrases (1-25)
0	JP Morgan expects that Scala will lower Nobel Biocare's growth forecast for 2007 from the current guidance of 23-25 pct, as well as the operating margin target from the current 34-35 pct.
1	JP Morgan predicts that Scala will reduce Nobel Biocare's 2007 growth forecast from 23-25 pct and their operating margin target from 34-35 pct.
2	According to JP Morgan, Scala is likely to cut Nobel Biocare's growth outlook for 2007 from the existing 23-25 pct and the operating margin goal from 34-35 pct.
3	JP Morgan anticipates that Scala will bring down Nobel Biocare's 2007 growth estimate from the projected 23-25 pct and the operating margin aim from 34-35 pct.
4	JP Morgan projects that Scala will decrease Nobel Biocare's growth forecast for 2007 from 23-25 pct to a lower figure, along with the operating margin from 34-35 pct.
5	JP Morgan foresees Scala reducing Nobel Biocare's 2007 growth forecast from 23-25 pct and their operating margin outlook from 34-35 pct.
6	JP Morgan suspects Scala will revise Nobel Biocare's 2007 growth projection downward from 23-25 pct and also lower the operating margin target from 34-35 pct.
7	JP Morgan plans for Scala to lower Nobel Biocare's growth forecast for 2007 from 23-25 pct and the operating margin target from 34-35 pct.
8	JP Morgan estimates that Scala will revise Nobel Biocare's 2007 growth expectations down from 23-25 pct and adjust the operating margin goal from 34-35 pct.
9	As per JP Morgan, Scala is anticipated to cut Nobel Biocare's forecasted 2007 growth from 23-25 pct and operating margin expectations from 34-35 pct.
10	JP Morgan opines that Scala will reduce Nobel Biocare's growth forecast for 2007 from the established 23-25 pct, along with the operating margin target from 34-35 pct.
11	JP Morgan suggests that Scala will lower Nobel Biocare's 2007 growth forecast from 23-25 pct and the operating margin aim from 34-35 pct.
12	According to JP Morgan's expectation, Scala will likely diminish Nobel Biocare's growth projection for 2007 from 23-25 pct and their margin goal from 34-35 pct.
13	JP Morgan indicates that Scala will decrease Nobel Biocare's growth forecast for 2007 from 23-25 pct and the operating margin target from 34-35 pct.
14	JP Morgan forecasts Scala will reduce Nobel Biocare's projected 2007 growth from 23-25 pct and the operating margin target from 34-35 pct.
15	JP Morgan's analysis shows that Scala will lower Nobel Biocare's 2007 growth forecast from 23-25 pct and the operating margin goal from 34-35 pct.
16	JP Morgan expects Scala to downgrade Nobel Biocare's 2007 growth forecast from the current 23-25 pct and also reduce the operating margin target from 34-35 pct.
17	JP Morgan believes Scala will scale back Nobel Biocare's 2007 growth prediction from 23-25 pct and curb the operating margin target from 34-35 pct.
18	JP Morgan has reported that Scala will lower the 2007 growth forecast for Nobel Biocare from 23-25 pct along with the operating margin from 34-35 pct.
19	JP Morgan's expectations are that Scala will cut Nobel Biocare's growth forecast for 2007 from 23-25 pct and also reduce the operating margin target from 34-35 pct.
20	JP Morgan envisages that Scala will lessen Nobel Biocare's 2007 growth forecast from 23-25 pct to a lower figure and the operating margin from 34-35 pct.
21	JP Morgan insights suggest that Scala will lower Nobel Biocare's projected 2007 growth from 23-25 pct and the operational margin aim from 34-35 pct.
22	JP Morgan is expecting Scala to downgrade Nobel Biocare's growth forecast for 2007 from the given 23-25 pct and the operating margin target from 34-35 pct.
23	JP Morgan anticipates Scala will decrease the growth forecast for Nobel Biocare's 2007 from the present 23-25 pct and their targeted operating margin from 34-35 pct.
24	JP Morgan sees Scala reducing Nobel Biocare's growth expectations for 2007 from 23-25 pct and similarly reducing the operating margin forecast from 34-35 pct.
25	JP Morgan's projections imply that Scala will lower Nobel Biocare's 2007 growth forecast from the current 23-25 pct and their operating margin target from 34-35 pct.

Bibliography

Abad Rocamora, Elias, Yongtao Wu, Fanghui Liu, Grigorios G Chrysos, and Volkan Cevher (2024). "Revisiting character-level adversarial attacks". In: *International Conference on Machine Learning (ICML)*.

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. (2023). "Gpt-4 technical report". In: *arXiv preprint arXiv: 2303.08774*.

AI4Finance-Foundation (2023a). *FinGPT: Open-Source Financial Large Language Models*. Accessed: 2024-05-15. URL: `https://github.com/AI4Finance-Foundation/FinGPT`.

— (2023b). *Making Data Notebook*. GitHub Repository. URL: `https://github.com/AI4Finance-Foundation/FinGPT/blob/master/fingpt/FinGPT_Sentiment_Analysis_v3/data/making_data.ipynb`.

Alzantot, Moustafa, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang (2018). "Generating natural language adversarial examples". In: *arXiv preprint arXiv: 1804.07998*.

Araci, Dogu (2019). "Finbert: Financial sentiment analysis with pre-trained language models". In: *arXiv preprint arXiv: 1908.10063*.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel

Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. (2018). "Universal sentence encoder". In: *arXiv preprint arXiv: 1803.11175*.

Cheng, Yong, Lu Jiang, and Wolfgang Macherey (2019). "Robust neural machine translation with doubly adversarial inputs". In: *arXiv preprint arXiv: 1906.02443*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv: 1810.04805*.

Ebrahimi, Javid, Anyi Rao, Daniel Lowd, and Dejing Dou (2017). "HotFlip: White-Box Adversarial Examples for Text Classification". In: *arXiv preprint arXiv: 1712.06751*. https://arxiv.org/abs/1712.06751.

Fatouros, Georgios, John Soldatos, Kalliopi Kouroumali, Georgios Makridis, and Dimosthenis Kyriazis (2023). "Transforming sentiment analysis in the financial domain with ChatGPT". In: *Machine Learning with Applications* 14, p. 100508.

Gao, Ji, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi (2018). "Black-box generation of adversarial text sequences to evade deep learning classifiers". In: *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, pp. 50–56.

Garg, Siddhant and Goutham Ramakrishnan (2020). "Bae: Bert-based adversarial examples for text classification". In: *arXiv preprint arXiv: 2004.01970*.

Goin, Michael (2022). *Classifying Finance Tweets in Real-Time with Sparse Transformers.* Accessed: 2023-02-06. URL: https://neuralmagic.com/blog/classifying-finance-tweets-in-real-time-with-sparse-transformers/.

Guo, Chuan, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela (2021). "Gradient-based adversarial attacks against text transformers". In: *arXiv preprint arXiv: 2104.13733.*

Iyyer, Mohit, John Wieting, Kevin Gimpel, and Luke Zettlemoyer (2018). "Adversarial Example Generation with Syntactically Controlled Paraphrase Networks". In: *Proceedings of NAACL-HLT*, pp. 1875–1885.

Jia, Robin and Percy Liang (2017). "Adversarial Examples for Evaluating Reading Comprehension Systems". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics.

Jin, Di, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits (Apr. 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* 34.05, pp. 8018–8025. DOI: 10.1609/aaai.v34i05.6311.

Leippold, Markus (2023). "Sentiment spin: Attacking financial sentiment with GPT-3". In: *Finance Research Letters* 55, p. 103957. ISSN: 1544-6123. DOI: 10.1016/j.frl.2023.103957.

Li, Jinfeng, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang (2018). "Textbugger: Generating adversarial text against real-world applications". In: *arXiv preprint arXiv: 1812.05271.*

Li, Linyang, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu (2020). "Bert-attack: Adversarial attack against bert using bert". In: *arXiv preprint arXiv: 2004.09984.*

Li, Xianming and Jing Li (2023). "Angle-optimized text embeddings". In: *arXiv preprint arXiv: 2309.12871.*

Loughran, Tim and Bill Mcdonald (2011). "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks". In: *The Journal of Finance* 66.1, pp. 35–65. DOI: https://doi.org/10.1111/j.1540-6261.2010.01625.x.

Maia, Macedo, 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*, pp. 1941–1942.

Malo, Pekka, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala (2014). "Good debt or bad debt: Detecting semantic orientations in economic texts". In: *Journal of the Association for Information Science and Technology* 65.4, pp. 782–796.

OpenAI (2024a). *Embeddings Guide.* Accessed: 2024-02-10. URL: https://platform.openai.com/docs/guides/embeddings.

— (2024b). *Hello GPT-4o.* Accessed: 2024-05-14. URL: https://openai.com/index/hello-gpt-4o/.

Papernot, Nicolas, Patrick McDaniel, Ananthram Swami, and Richard Harang (2016). "Crafting adversarial input sequences for recurrent neural networks". In: *MILCOM 2016-2016 IEEE Military Communications Conference.* IEEE, pp. 49–54.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Samanta, Suranjana and Sameep Mehta (2017). "Towards crafting text adversarial samples". In: *arXiv preprint arXiv: 1707.02812.*

Sinha, Ankur, Satishwar Kedas, Rishu Kumar, and Pekka Malo (2022). "SEntFiN 1.0: Entity-aware sentiment analysis for financial news". In: *Journal of the Association for Information Science and Technology* 73.9, pp. 1314–1335.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. (2023). "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv: 2307.09288.*

Usynin, Dmitrii, Alexander Ziller, Marcus Makowski, Rickmer Braren, Daniel Rueckert, Ben Glocker, Georgios Kaissis, and Jonathan Passerat-Palmbach (2021). "Adversarial interference and its mitigations in privacy-preserving collaborative machine learning". In: *Nature Machine Intelligence*

3.9, pp. 749–758. DOI: 10.1038/s42256-021-00390-3. URL: https://doi.org/10.1038/s42256-021-00390-3.

Wang, Boxin, Hengzhi Pei, Boyuan Pan, Qian Chen, Shuohang Wang, and Bo Li (2020). "T3: Tree-Autoencoder Constrained Adversarial Text Generation for Targeted Attack". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6134–6150.

Wang, Oliver (2023). *News with GPT Instructions*. Hugging Face Dataset. URL: https://huggingface.co/datasets/oliverwang15/news_with_gpt_instructions.

Wang, Zimu and Hong-Seng Gan (2023). "Multi-level adversarial training for stock sentiment prediction". In: *2023 IEEE 3rd International Conference on Computer Communication and Artificial Intelligence (CCAI)*. IEEE, pp. 127–134.

Wieting, John and Kevin Gimpel (2018). "ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Wu, Shijie, 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". In: *arXiv preprint arXiv: 2303.17564*.

Xie, Yong and Dakuo Wang (2022). "A Word is Worth A Thousand Dollars: Adversarial Attack on Tweets Fools Stock Prediction". In: *Conference of the North American Chapter of the Association for Computational Linguistics*.

Yang, Hongyang, Xiao-Yang Liu, and Christina Dan Wang (2023). "Fingpt: Open-source financial large language models". In: *arXiv preprint arXiv: 2306.06031*.

Yang, Shuo, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica (2023). "Rethinking benchmark and contamination for language models with rephrased samples". In: *arXiv preprint arXiv: 2311.04850*.

Zhao, Zhengli, Dheeru Dua, and Sameer Singh (2018). "Generating Natural Adversarial Examples". In: *International Conference on Learning Representations*.

Zou, Andy, Zifan Wang, J Zico Kolter, and Matt Fredrikson (2023). "Universal and transferable adversarial attacks on aligned language models". In: *arXiv preprint arXiv: 2307.15043*.