NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001311

# SCUNLP-1 at the NTCIR-17 FinArg-1 Task: Enhancing Classification Prediction through Feature Generation Based on ChatGPT

Ya-Mien Cheng
Dept. of Data Science
Soochow University
Taiwan
mian12080916@gmail.com

Jheng-Long Wu
Dept. of Data Science
Soochow University
Taiwan
jlwu@gm.scu.edu.tw

## ABSTRACT

This paper will provide a comprehensive overview of our participation in two shared tasks of NTCIR17 FinArg-1: Argument-based Sentiment Analysis (Earning Calls), including Argument Unit Classification (ECCAC) and Argument Relation Detection and Classification (ECCAR). We submitted three rounds of predictions for each task during the final evaluation. The study was to determine whether a sentence acted as a premise and predict its sentiment as none, support, or attack. Our proposed solution involves gathering insights through conversations with large language models. In both tasks, we categorized questions into two distinct types: those directly addressing classification categories and those about assessing financial market relevance. These questions were approached from three perspectives: the general public, professional financial market investors, or without specifying a particular view. These insights are then integrated into the features for model prediction. The experimentation mainly consisted of two scenarios: using only the original data and employing both the original data and ChatGPT's answers during training. Ultimately, we discovered that incorporating ChatGPT's responses alongside the original data yielded the highest scores in both tasks, surpassing other experimental configurations that relied solely on either the original data or ChatGPT alone. In the task of ECCAC, a 71.07% Macro-F1 was attained, while ECCAR yielded a score of 54.60% in Macro-F1. Although the performance in the ECCAC task did not significantly surpass other versions, it remained the most successful among the submitted versions.

## KEYWORDS

Data Augmentation, Prompt, ChatGPT, Feature Generation

## TEAM NAME

SCUNLP-1

## SUBTASKS

FinArg-1: Fine-grained Argument Understanding in Financial Analysis (Chinese)

Subtask 1: Argument Unit Classification

Subtask 2: Argument Relation Detection and Classification

## 1 INTRODUCTION

The SCUNLP-1 team participated in the Argument-based Sentiment Analysis (Earnings Call), which has two subtasks, including Argument Unit Classification (ECCAC) and Argument Relation Detection and Classification (ECCAR) [5]. In the two tasks in this bullish claim. The dataset in the first subtask includes two classes, which means it is a binary classification: premises and claims. Further sentiment labels are assigned to these arguments. The term "premise" refers to the antecedent conditions necessary for the emergence or development of a phenomenon. The term "claim" refers to a publicly declared assertion or announcement representing a fact already occurring. The other dataset has three classes for multiclass classification: none, support, and attack [1]. The primary objective is to conduct sentiment analysis on sentences to discern the stance of the discourse.

In previous text sentence categorization tasks, extracting new features through techniques such as tokenization or summarization has been common practice. However, these methods may yield less domain-specific content and knowledge.

Recently, large-scale chat-based language models have gained popularity, allowing models to adopt different roles by setting prompts. Therefore, we propose using conversations with a large language model (LLM) enriched with domain-specific content to generate responses and experiment with whether this approach can effectively enhance text classification performance. This article presents a method for developing new features to improve training effectiveness based on conversations with an LLM for question-answering, GPT-3.5 turbo, as proposed by OpenAI in recent years. We engage in prompts with the LLM question-answering model using specific conversation templates and utilize the responses as new features, attempting to merge them with the existing features.

Our primary objective is to leverage generative pre-trained transformers (GPT) to generate additional information and labels for arguments extracted from the Earnings Call dataset. To achieve this, we provide multifaceted insights into the issue from various perspectives and incorporate auxiliary information for label classification using different models. This augmentation enriches the feature set of the data, enabling a more detailed and in-depth analysis of investor sentiment and discourse dynamics in the financial domain.

In the subsequent sections, we delve into the details of our methodology, the unique role of GPT in improving argument analysis, and the implications of our research for the broader field of financial analysis and sentiment assessment. Ultimately, our study contributes to a deeper comprehension of argument mining in finance, shedding light on the complex interplay between emotions, arguments, and financial decision-making.

# 2 RELATED WORK

## 2.1 Text Data Augmentation

Text data augmentation is a technique used to increase the diversity and size of a text dataset by creating new variations of existing text while preserving the original meaning. This is particularly valuable for training natural language processing (NLP) models, as more extensive and more varied datasets often lead to better model performance [3]. Prior research has introduced NLP data augmentation techniques. Several text augmentation techniques based on synonym replacement have been proposed so far. However, these techniques have a standard limitation in that there is a possibility of generating incorrect text from the content of an image when replacing the synonym for a noun word [9]. Through language translation, it is possible for different words to represent the same meaning, a phenomenon known as back-translation. However, the scalability of augmented corpora is constrained since this approach involves substituting existing sentence elements [7]. A technique known as Contextual Augmentation can address the challenges above. The research is based on the assumption that sentences exhibit invariance, meaning that even when words within sentences are replaced with other words sharing morphological relationships, the sentences continue to convey conceptually similar ideas. This process stochastically replaces words at various positions within the sentence with alternative words predicted by a bidirectional language model. While numerous words can be predicted based on the context, they are appropriately utilized to augment the original words [10].

## 2.2 Prompt Engineering

Prompt engineering is a skill set needed to converse effectively with LLMs. Prompts are instructions given to an LLM to enforce rules, automate processes, and ensure specific qualities (and quantities) of generated output [17], which is considered a control variable for modulating the output distribution of the LLM [4]. Previous studies have traditionally classified prompts into two main categories: prompts and continuous prompts. Discrete prompts are well-suited for addressing concise questions as they often involve posing individual questions separately. However, this approach necessitates restructuring the question's format, which can decrease the relevance between the questions and the original target information [12].

On the contrary, continuous prompts find more applicability in longer, more complex questions. This is because they enable some gradient information propagation between the questions. Yet, this increased complexity in computation is a known issue, and adjustments are often required in practice [11]. Moreover, some research has confirmed that the format of prompts significantly influences the responses generated by GPT models [2]. Adopting a Chain-of-Prompting approach has been shown to enhance the logical reasoning capabilities of these models [16], which means a series of intermediate reasoning steps.

## 2.3 ChatGPT

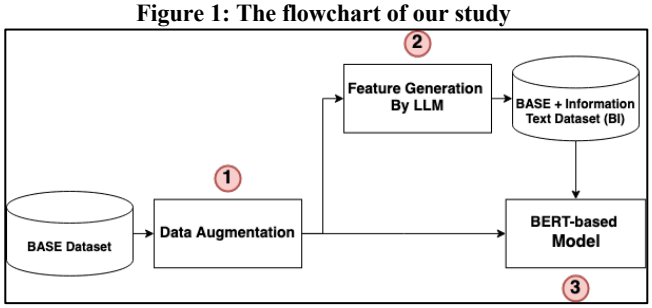
The field of natural language processing has witnessed substantial advancements in recent years. These transformer models, such as BERT [6] or GPTs [13], have rapidly become the dominant architecture for natural language processing, surpassing alternative neural models, such as convolutional and recurrent neural networks, in performance for tasks in both natural language understanding and natural language generation [18]. These research endeavors have demonstrated remarkable success in a range of applications, including but not limited to speech recognition, part-of-speech tagging, and syntactic parsing. Recently, an increasing number of researchers have applied this approach to the field of Information Retrieval, thereby augmenting the effectiveness of information retrieval processes [15]. The generation of new auxiliary information by LLMs has previously faced skepticism regarding the distinction between "new information" and the extension of existing knowledge through known information. Recent research has emerged to compare the semantic concepts generated by GPT-3 with the extended ideas recognized by humans, verifying that it can accurately reflect essential aspects of human conceptual knowledge [8].

# 3 METHODS

This section will describe how two subtasks (ECCAC and EC-CAR) generate new features from the original data and incorporate these new features back into the original dataset for subsequent classification predictions. Figure 1 depicts the workflow diagram for this research study. To begin with, during the data exploration phase, it was observed that the dataset for ECCAR involves sentiment classification based on two distinct sentences, unlike the ECCAC task, which permits more straightforward downstream processing. Consequently, a preliminary step was taken to merge individual sentences for further analysis. Next, recognizing the imbalance in the data for both tasks, a method was employed to generate similar but different sentences, an oversampling technique, thereby equalizing the quantity of each class.

To ensure a consistent derivation of response templates from the Language Model (LM), a structured framework of dialogic queries was developed for LLM input. The primary objective was to solicit responses that encompass emotional context and convey the discernment of a statement's conjectural nature. Subsequently, the generated responses were seamlessly integrated with the source sentences. The ensuing data preprocessing steps encompassed removing extraneous characters and text optimization to render it suitable for input into the final model.

**Figure 1: The flowchart of our study**




## 3.1 Data Augmentation

Oversampling and under-sampling can achieve a balanced data distribution across all categories. Considering the low proportion of data in the two types for the unit classification task, under-sampling would result in an insufficient dataset for improved classification performances. Therefore, the Oversampling method is adopted. The oversampling of the data generation methodology employed in the study is contextual augmentation. This technique enables the model to augment sentences without compromising label compatibility, which is one of the reasons for its selection. Since sentences are inherently associated with specific categories, this approach allows for word replacements within sentences without affecting their categorization. This process can be extended to consecutive sentences, where similar words are substituted, ensuring the sentences remain contextually coherent. This method serves as a means of data augmentation, addressing the severe class imbalance present in our dataset by generating the necessary volume of data. Fig. 2 illustrates an example of sentence expansion using Contextual Augmentation.

**Figure 2: An example of techniques for contextual augmentation to perform word substitution**

ORIGINAL:
we added apple pay we added apple music sep we have stated that we want to double the size of the services business from the level that we had in fiscal 2016 by 2020.
==========
GENERATED:
they add mac pay we add iphone music sep we have stated that what you to two same larger of that clients businesses from the level that we had in fiscal 2017 another 2019.

Original sentence and generated sentences have same context but differing slightly. For example, a sentence originally ending with "... in fiscal 2016 by 2020" is substituted with "... in fiscal 2017 by 2019". Firstly, all the required sentences for generation are merged into one. Since generating individual sentences and paragraphs is the same. Then, not all data from the dataset needs to be generated. We randomly extract the required quantity for generation. Next, the desired generation category needs to be specified. Contextual Augmentation uses existing sentences in that category as references to identify similar aspects in words. The Contextual Augmentation technique involves the utilization of a bidirectional LM based on the provided type. It extends sentences before and after the target sentence and replaces words with semantically similar alternatives. This process ensures an equal data distribution across all categories for the two classification tasks, as demonstrated in Tables 1 and 2.

## 3.2 Feature Generation by LLM

During the model training phase, the quality of the training data is of paramount importance. This holds particularly true in textual data, where each entry in the dataset must adhere to a consistent format and content. Such uniformity ensures that descriptions are appropriately aligned with the relevant aspects of the task. The primary focus of our research is to investigate whether text generated by large language models can aid in text classification. We devised a strategy to utilize GPT to generate features to accomplish this. The design of prompts to guide the content generated by GPT is essential to our approach. When formulating these templates, our attention is not solely directed toward formatting but also toward explicitly conveying the context of the question and the expected response. The question templates for the two classification tasks are illustrated separately in Fig. 3 and 8. In devising the prompts, considering the financial domain's relevance to both unit classification and relation detection and classification tasks, we formulated questions that involved financial subject matters. The "<Post>" section of the form should be replaced with each specific data point provided in the dataset.

### 3.2.1 Prompt Design for ECCAC

In this task, we have designed three distinct perspectives for examining the problem, as shown in Fig. 3: financial domain professionals (P), the general public (G), and unspecified individuals (N). Furthermore, we have devised two types of questions, namely, those directly addressing the task's classification labels "Premise/Claim" (P) and those inquiring about financial-related issues (F).

**Figure 3: The demonstration of the ECCAC task prompt**

Sentence: < Post >
Based on the sentence above, please answer the questions individually and provide the responses in the following format below:
<PPQ>
<GPQ>
<NPQ>
<NFQ>

- <PPQ>: This question is designed for financial domain professionals, as shown in Fig. 4. It directly queries the dataset's concepts corresponding to "Premise/Claim." While it does not explicitly request information about "Premise/Claim," the wording of the statements also encompasses interchangeable terms with similar concepts.

**Figure 4: A template designed for PPQ questions**

Question 1:
1. For investors who have been involved in the financial sector for many years, l clarify this sentence contains more (factual/hypothetical) elements.
2. For investors who have been involved in the financial sector for many years, I specify this sentence is (subjective/objective).

- <GPQ>: This question is designed for the general public, as shown in Fig. 5. It directly queries the dataset's concepts corresponding to "Premise/Claim." While it does not

explicitly request information about "Premise/Claim," the wording of the statements also encompasses interchangeable terms with similar concepts.

**Figure 5: A template designed for GPQ questions**

Question 2:
1. From the perspective of the general public, I clarify this sentence contains more (factual/hypothetical) elements.
2. From the perspective of the general public, I specify this sentence is (subjective/objective).

- <NPQ>: This question is designed for unspecified individuals, as shown in Fig. 6. It directly queries the dataset's concepts corresponding to "Premise/Claim." While it does not explicitly request information about "Premise/Claim," the wording of the statements also encompasses interchangeable terms with similar concepts.

**Figure 6: A template designed for NPQ questions**

Question 3:
1. This sentence (is/isn't) discussing a review of strategies.
2. This statement is considered (positive/negative/neutral).

- <NFQ>: This question is designed for unspecified individuals, as shown in Fig. 7. Financial domain professionals. The questions primarily revolve around issues related to the financial domain, including the future and past developments of the stock market. The main focus of the discussion is on stocks, funds, and similar topics.

**Figure 7: A template designed for NFQ questions**

Question 4:
1. This sentence (is/isn't) about adjusting-future strategies.
2. This sentence is discussing (company/industry/forex stocks/market changes/other) topics.
3. This sentence (is/isn't) about planning for development.
4. This sentence (is/isn't) discussing past developments.

### 3.2.2 Prompt Design for ECCAR

In this task, three distinct perspectives have structured to approach the problem, as shown in Fig 8: financial domain professionals (P), the general public (G), and unspecified individuals (N). Two questions have been designed, directly targeting the task's sentiment label classification (S), encompassing three tones – none, support, and attack. Additionally, questions are inquiring about financial-related issues (F).

**Figure 8: The demonstration of the ECCAR task prompt**

Sentence 1: < Post 1 >
Sentence 2: < Post 2 >
Based on the two sentences above, please answer the questions individually and provide the responses in the following format below:
<GSQ> <GFQ> <PSQ> <PFQ>

- <PSQ>: This question is designed for financial domain professionals, as shown in Fig. 9. It directly queries the emotional stance of the none, support, or attack.

**Figure 9: A template designed for PSQ questions**

Question A:
1. From the perspective of an investor who has been involved in the financial field for many years this statement is considered (positive/negative/neutral).

- <GSQ>: This question is designed for the general public, as shown in Fig. 10. It directly queries the emotional stance of the none, support, or attack.

**Figure 10: A template designed for GSQ questions**

Question B:
1. From the perspective of the general public, this statement is considered (positive/negative/neutral).

- <NSQ>: This question is designed for unspecified individuals, as shown in Fig. 11. It directly queries the emotional stance of the none, support, or attack.

**Figure 11: A template designed for NSQ questions**

Question C:
1. This sentence (/not) include the concept of 'market decline, investor pessimism, and widespread price decline'.
2. This sentence (/not) include the concept of 'market ascent, investor optimism, and widespread price increase'.

- <PFQ>: This question is designed for financial domain professionals, as shown in Fig. 12. These questions pertain to financial topics, explicitly concerning future market developments, whether positive or negative.

**Figure 12: A template designed for PFQ questions**

Question D:
1. From the perspective of an investor who has been involved in the financial field for many years, this statement is currently leaning towards a (bear market/bull-market).
2. From the perspective of an investor who has been involved in the financial field for many years, there (will/won't) be a trading halt in the near future.
3. From the perspective of an investor who has been involved in the financial field for many years, there (will/won't) be a circuit breaker halt in the near future.

- <GFQ>: This question is designed for the general public, as shown in Fig 13. These questions pertain to financial topics, explicitly concerning future market developments, whether positive or negative.

**Figure 13: A template designed for GFQ questions**

Question E:
1. From the perspective of the general public, this statement is currently leaning towards a (bear market/bull market).
2. From the perspective of the general public, there (will/won't) be a trading halt in the near future.
3. From the perspective of the general public, there (will/won't) be a circuit breaker halt in the near future.

While the majority of responses conform to the expected format, there is a small subset that deviates from these expectations. These responses may consist of only a few key terms, such as responding with a simple "yes" or "no," or they may include additional introductory and concluding paragraphs. Incidents of such non-conforming responses are relatively infrequent, typically comprising fewer than ten instances within each dataset. As a result, these responses have not been subjected to any special treatment and have been treated as standard data entries for subsequent model training. Following the retrieval of responses from the GPT model for various types of questions, these responses, being complete sentences, are directly appended to the original dataset as new data content.

## 3.3 BERT-Based Model

Given that the datasets exclusively contained English text, English pre-trained models were selected for text classification. The Hugging Face community offers an array of finely tuned BERT-based models. Several domain-agnostic models were established, such as "*distilbert-base-multilingual-cased*" [14] and "*distilbert-base-uncased- finetuned-sst-1-2-english*" [14], in addition to models tailored to the financial domain. Significantly, the "*yiyanghkust/finbert-tone*" [19] model consistently demonstrated exceptional performance across various training versions. Consequently, it was adopted to establish a submission record for the ECCAC task. Furthermore, models specialized in sentiment analysis were integrated into the research. The "*nlptown/bert-base-multilingual-uncased-sentiment*" [6] and "*bhadresh-savani/distilbert-base-uncased-emotion*" [14] model delivered noteworthy results and were applied to construct a submission record for the ECCAR task. Two experimental versions were conducted. One involved training with the original dataset (BASE), while the other incorporated the original dataset and the Information text dataset (BI). The features comprised descriptions of individual sentences within the dataset, with the target being the corresponding classification label. These labels corresponded to Premise/Claim for ECCAC, whereas for ECCAR, they were support, attack, and none. During the training of the BERT models, it is widely acknowledged that there is a limitation on the maximum sequence length, typically capped at 512 tokens.

Nevertheless, post-data preprocessing, which involved the removal of stop words, punctuation, and the uniform conversion of text to lowercase, was observed that instances of data exceeding the 512-token limit were relatively infrequent. Hence, direct training was deemed feasible.

**Table 1: The distribution of the ECCAC dataset**

| Dataset | Label | Train | Dev | Test | All (%) |
|---|---|---|---|---|---|
| BASE | Premise | 4062 | 508 | 508 | 52 |
| | Claim | 3691 | 461 | 461 | 48 |
| | All | 7753 | 969 | 969 | 100 |
| BI | Premise | 4062 | 508 | 508 | 52 |
| | Claim | 4062 | 508 | 508 | 52 |
| | All | 8124 | 1016 | 1016 | 100 |

**Table 2: The distribution of the ECCAR dataset**

| Dataset | Label | Train | Dev | Test | All (%) |
|---|---|---|---|---|---|
| BASE | None | 1600 | 200 | 200 | 29 |
| | Support | 3859 | 482 | 482 | 70 |
| | Attack | 62 | 8 | 8 | 1 |
| | All | 5521 | 690 | 690 | 100 |
| BI | None | 3859 | 482 | 482 | 33.33 |
| | Support | 3859 | 482 | 482 | 33.33 |
| | Attack | 3859 | 482 | 482 | 33.33 |
| | All | 11577 | 1446 | 1446 | 100 |

# 4 EXPERIMENTS

In this section, we will primarily present the model experimental results of our study. We submitted three sets of prediction results for each task, totaling six, as indicated in Tables 3 and 8. As mentioned in the previous section, the model selection process is characterized by the fine-tuned models we ultimately submitted. However, in addition to these models, we also experimented with several others, including "*roberta-large-mnli*" and "*bert-base-uncased*," among others. Ultimately, these alternative models did not perform as well as expected, and thus, they were not submitted for evaluation.

## 4.1 Submission of Results

According to the task of FinArg-1, the primary metric of reference was macro-F1. The three best-performing submissions had varying learning rates, models, and datasets in the ECCAC task, as shown in Table 3. However, it's noteworthy that these learning rates of submissions and models were chosen based on the best-performing model parameters. Consequently, these submissions outperformed versions that used GPT-generated supplementary information compared to those that relied solely on the original dataset for predictions. However, in the ECCAC task, the difference in performance was minimal, with a macro-F1 of 71.07 compared to 70.53, showing a marginal improvement. In contrast, the difference was more significant in the ECCAR task, with a macro-F1 score of 54.6 compared to 49.00, and it was one of the most outstanding among all the submitted versions, as shown in Table 4. Based on the outcomes of these two tasks, when adding supplementary information to the data using the GPT approach, we can reasonably infer that the problem design for the ECCAC task could have been more effective. This suggests that the actual helpfulness of the additional information could have been improved, resulting in an improvement that, while present, was not statistically significant. It's also evident that in the ECCAC task, models fine-tuned for unrelated or specific-task-related domains are more suitable due to the broad categorization of classes. However, for the ECCAR task, models fine-tuned for sentiment analysis, which is more relevant to the task, yield better results. As a result, non-domain-specific fine-tuned models, when compared to other versions, didn't perform significantly, leading to their non-selection for submission.

**Table 3: Submission results for NTCIR-17 FinArg-1 ECCAC Task**

| Submission name | Model | Dataset | Learning rate | Micro-F1 | Macro-F1 | Weight-F1 | Rank |
|---|---|---|---|---|---|---|---|
| SCUNLP-1-2 | *distilbert-base-multilingual-cased* | BI | 2e-05 | 71.10% | 71.07% | 71.02% | 17 |
| SCUNLP-1-3 | *distilbert-base-uncased-finetuned-sst-1-2-english* | BASE | 4e-05 | 71.10% | 70.53% | 70.73% | 18 |
| SCUNLP-1-1 | *yiyanghkust/finbert-tone* | BASE | 5e-05 | 68.73% | 68.62% | 68.53% | 19 |

**Table 4: Submission results for NTCIR-17 FinArg-1 ECCAR Task**

| Submission name | Model | Dataset | Learning rate | Micro-F1 | Macro-F1 | Weight-F1 | Rank |
|---|---|---|---|---|---|---|---|
| SCUNLP-1-3 | *bhadresh-savani/distilbert-base-uncased-emotion* | BI | 4e-05 | 72.17% | 54.60% | 72.35% | 8 |
| SCUNLP-1-2 | *bhadresh-savani/distilbert-base-uncased-emotion* | BASE | 4e-05 | 68.55% | 49.00% | 68.57% | 16 |
| SCUNLP-1-1 | *nlptown/bert-base-multilingual-uncased-sentiment* | BASE | 2e-05 | 68.70% | 45.68% | 68.05% | 18 |

## 4.2 Analysis of Prediction Results

Subsequently, the results of the predictions for both tasks were examined. This analysis provided insights into the model's performance across different categories, allowing for targeted improvements in cases where predictions were less than ideal. It's important to note that this section focuses solely on the best submission versions for both tasks.

### 4.2.1 Analysis for ECCAC

The correct answers and prediction results for this task are presented as a confusion matrix, as shown in Fig. 14. The category "Premise," when the actual label, is more prone to misclassification, with an error rate of approximately 35.4%. In contrast, the error rate for the actual label "Claim" is about 21.7%.

Next, we investigated whether sentence length affects the model's prediction error rate. This analysis is presented as a histogram, as shown in Fig. 15. The red bars represent incorrect predictions, while the green bars represent correct ones. The x-axis indicates the length of the data sentences, and the y-axis represents the quantity. From Fig. 15, we observe that sentence length moderately impacts prediction accuracy for this task. While the correlation could be higher, there is still an effect. Notably, prediction errors in sentence lengths between 350 and 400 are more elevated than correct predictions. For shorter sentences (length between 1 to 50), this may be due to the limited informational content of the sentences, leading to a relatively higher error rate compared to other categories.

### 4.2.2 Analysis for ECCAC

The correct answers for this task and the predicted results are presented in the form of a confusion matrix, as shown in Fig. 16. It can be observed that the "Attack" category is more prone to misclassification during predictions, with an error rate of approximately 75%. There is a high probability, as high as 62%, that it will be predicted as "Support." However, it's important to note that the original dataset contains very few instances of the "Attack" class, which causes the model to lean toward predicting the majority class, "Support."

Even though we have employed Data Augmentation techniques, the improvement still needs to be improved. Interestingly, data from the "None" category has a 42% probability of being predicted as "Support." This suggests that discerning the stance of "None" might be more challenging compared to "Attack" or "Support" due to its less distinct nature.

Subsequently, we investigated whether sentence length affects the model's prediction error rate. This is visualized in a histogram, as shown in Fig. 17, where the red bars represent prediction errors, and the green bars represent correct predictions. The x-axis represents the length of data sentences, and the y-axis represents the quantity. Predictions for sentences over 400 characters are notably more accurate than shorter ones. Given that this task involves judging the sentiment stance of two sentences, it naturally involves longer data sentences than another task. We reasonably infer that the decline in accuracy due to length is related to BERT's token limit. Shortening sentences by summarizing them should be considered a potential approach for future optimization.

**Figure 14: Bar chart of prediction and answer for the ECCAC Task**

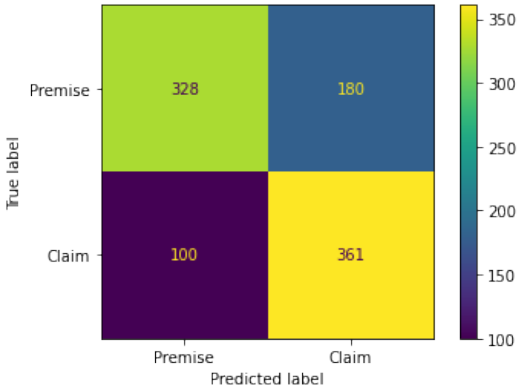


**Figure 15: Histogram of correct and incorrect prediction sentence lengths for the ECCAC Task**

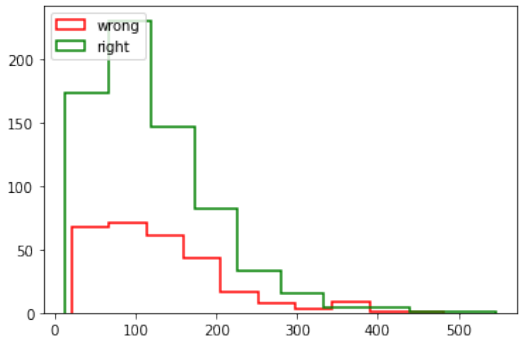



**Figure 16: Confusion matrix of prediction and answer for the ECCAR Task**

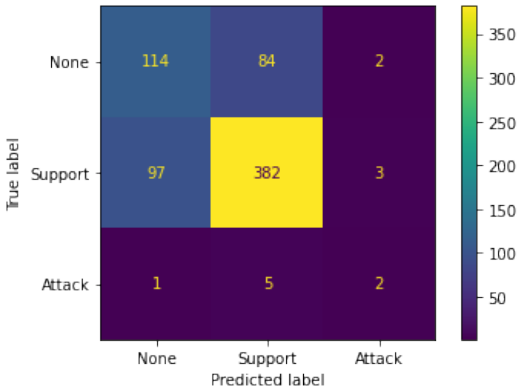
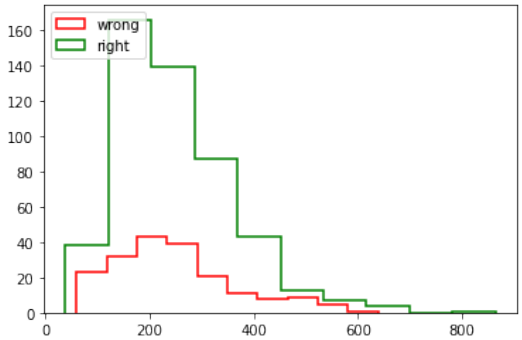



**Figure 17: Histogram of correct and incorrect prediction sentence lengths for the ECCAR Task**




## 5 CONCLUSIONS

In this study, we employed the Contextual Augmentation method to address the issue of extreme class imbalance within the text dataset. Additionally, we designed prompts tailored to the financial domain, leveraging the GPT-3.5 turbo model to enhance the information content of the data. Finally, we utilized a BERT-based model and identified pre-existing fine-tuned models within finance and sentiment analysis for predictive tasks.

Furthermore, we explored the potential of GPT-3.5 Turbo, an LLM explicitly designed for generative tasks. We developed a series of prompts that could enhance the informativeness of the generated text. These prompts included questions about financial market developments and inquiries about the emotional stance within sentences. Additionally, we experimented with various perspectives from which these questions were posed. Our experimental results demonstrated that incorporating information generated by GPT could enhance model performance.

It's essential to note that the prompts we designed represent only one approach. We have not yet explored methods involving information extraction using GPT, and not all prompt designs may be equally effective in improving model performance. Experimenting with inquiries from different domains or utilizing diverse questioning strategies may yield more effective outcomes. This presents a topic for future investigations.

## REFERENCES

[1] Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022. It's Time to Reason: Annotating Argumentation Structures in Financial Earnings Calls: The FinArg Dataset. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. 163–169.

[2] Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask me anything: A simple strategy for prompting language models. arXiv preprint arXiv:2210.02441 (2022).

[3] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *Comput. Surveys* 55, 7 (2022), 1–39.

[4] Aman Bhargava, Cameron Witkowski, Manav Shah, and Matt Thomson. 2023. What's the Magic Word? A Control Theory of LLM Prompting. arXiv preprint arXiv:2310.04444 (2023).

[5] Chung-Chi Chen, Chin-Yi Lin, Chr-Jr Chiu, Hen-Hsen Huang, Alaa Alhamzeh, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2023. Overview of the NTCIR- 17 FinArg-1 Task: Fine-grained argument understanding in financial analysis. In *Proceedings of the NTCIR-17 Conference*.

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

[7] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381* (2018).

[8] Hannes Hansen and Martin N Hebart. 2022. Semantic features of object concepts generated with GPT-3. *arXiv preprint arXiv:2202.03753* (2022).

[9] Museong Kim and Namgyu Kim. 2021. Text Augmentation Using Hierarchy-based Word Replacement. *Journal of The Korea Society of Computer and Information* 26, 1 (2021), 57–67.

[10] Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201* (2018).

[11] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. GPT understands, too. *AI Open* (2023).

[12] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773* (2021).

[13] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[14] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[15] Fei Song and W Bruce Croft. 1999. A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management*. 316–321.

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[17] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* (2023).

[18] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.

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