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/10.20736/0002001297

IMNTPU at the NTCIR-17 FinArg-1 Argument-based Sentiment Analysis and Identifying Attack and Support Argumentative Relations in Social Media Discussion Threads

Chia-Tung Tsai
Information Management,
National Taipei University
New Taipei City, Taiwan
s711136106@gm.ntpu.edu.tw

Wen-Hsuan Liao
Information Management,
National Taipei University
New Taipei City, Taiwan
s711136109@gm.ntpu.edu.tw

Hsiao-Chuan Liu
Information Management,
National Taipei University
New Taipei City, Taiwan
s711136108@gm.ntpu.edu.tw

Vidhya Nataraj
Smart Healthcare Management,
National Taipei University
New Taipei City, Taiwan
vidhyanataraj99@gmail.com

Tzu-Yu Liu
Business Administration,
National Taipei University
New Taipei City, Taiwan
s411084008@gm.ntpu.edu.tw

Mike Tian-Jian Jiang
Zeal Co., Ltd
Tokyo, Japan
tmjiang@gmail.com

Min-Yuh Day*
Information Management,
National Taipei University
New Taipei City, Taiwan
myday@gm.ntpu.edu.tw

ABSTRACT
In recent years, there has been a surge of interest in argument-based sentiment analysis and the identification of argumentative relationships in social media. These tasks encompass sentiment analysis of premises and claims, as well as the classification of argumentative relationships. Within these tasks, we have developed a fine-tuning method for transformer models. To evaluate and showcase this concept, we established a comprehensive framework to test and display the performance of BERT, RoBERTa, FinBERT, ALBERT, and GPT 3.5-turbo models on financial data and social media texts. Ultimately, the experimental results of these sub-tasks validate the effectiveness of our strategies. The primary contribution of our research is our proposal of two key elements: fine-tuning predominantly with BERT models and employing GPT for generative classification, aiming to enhance the identification of argumentative classifications. Through fine-tuning techniques, the state-of-the-art models can achieve better performance than the baseline.

KEYWORDS
Sentiment Analysis, Social Media, Argument Mining, Fine-Tuning, Transformers

TEAM NAME
IMNTPU

SUBTASKS
Task-1: Argument-based Sentiment Analysis

(Earnings Conference Call)

Subtask 1: Argument Unit Classification

Subtask 1: Argument Relation Detection and Classification

Task-2: Identifying Attack and Support Argumentative Relations in Social Media Discussion Threads (Social Media)

1. INTRODUCTION
Recent developments in FinTech have heightened the need for the task of natural language processing. The accessibility of textual content online—such as financial articles, earnings conference calls that have been shown on social media—provides plenty of resources for technological application in financial domain [1].

Sentiment analysis was a common example for the tasks of natural language processing. Advanced sentiment analysis techniques, specifically designed for financial data, categorized sentiments into positive, negative, and neutral labels. Through the application of machine learning algorithms tailored for financial sentiment extraction, insights into market trajectories were derived. With the application of sentiment analysis, it not only provided clarity on current market sentiments but also demonstrated predictive capabilities for forecasting market trends and directions. The accumulated findings from this analysis hold significant potential in refining investment strategies and enhancing decision-making processes within the FinTech sector.

With the emergence of Transformer-based pretrained models, research on large language models (LLMs) has expanded significantly. In addition, the applications of large language models have become increasingly diverse. Utilizing datasets sourced from such as social media, FinTech platforms and traditional financial databases, the perceptions of both managers and investors regarding market dynamics were analyzed.

According to the reasons above, the research objective is employing diverse computational models for the sentiment analysis of data derived from earnings conference calls and social media. By doing so, we aim to assess the emotional sentiment embedded within the input data. Following the sentiment evaluation, a

comparative analysis will be undertaken to ascertain the accuracy and efficacy of each model. This endeavor is pivotal to enhancing our understanding of how different models perform in the realm of financial data sentiment assessment.

In the subsequent portions of this document, Section 2 provided an overview of the related work to this study. Section 3 explicated the research methodology employed. This was followed by an examination of the experimental data in Section 4. The conclusion of this research was shown in Section 5.

2. RELATED WORK

In this section, we reviewed the previous work with the relevant content in FinArg-1. We introduced application of sentiment analysis in social media and the earning conference calls domain. At last, we showed the related work on Transformer-based models.

2.1 Social Media

Social media is a usual way to analyze and classify people's emotions and opinion, and sentiment analysis makes good use of NLP techniques to do this task [2]. Sentiment analysis is based on detect the polarity of text and classify them on this premise. A common sentiment polarity is categorized as positive, negative or neutral [3].

In the past, a lot of academic research use LSTM, and CNN structure to sentiment analysis in social media [4]. With the emergence of Transformer-based pretrained models, research on large language models (LLMs) has expanded significantly. Therefore, there are more and more researchers use LLM to this application.

Earnings conference calls can loosely encompass two parts: firstly, a concise presentation detailing the company's performance for the previous quarter and forecasts for the upcoming period; and secondly, a subsequent Q&A session, also named as discussion portion [5].

By examining the text and audio meeting minutes, which was included the dictation from managers during earnings conference calls, we can discern hidden insights and previously unrecognized information in the market [6]. Therefore, it is very important to explore and analyze the data of earning conference calls.

2.2 Transformer-based models

Transformer architecture was introduced in 2017, not only bringing about a technological revolution but also offering unprecedented possibilities for Natural Language Processing (NLP) [7]. Before its advent, recurrent models like RNNs and LSTMs were widely favored. However, the emergence of the Transformer challenged their dominance in sequential data analysis.

What distinguishes the Transformer is its "self-attention" mechanism. This allows the model to consider every part of the input sequence simultaneously, capturing context and relationships between elements more effectively. This design enables it to encompass an entire sentence in a single computation, proving more efficient and accurate than the step-by-step calculation of recurrent models. Moreover, it incorporates techniques like multi-head attention, positional encoding, and feed-forward networks, enhancing its capabilities in language comprehension.

Models based on the Transformer, such as BERT, GPT, and T5, have gradually become the new standard in NLP. They predominantly utilize the Transformer's self-attention mechanism, conducting deep training on large datasets, delving into the intricacies and structures of language. Specifically, BERT is renowned for its bidirectional training strategy, deeply probing textual contexts [8]. GPT employs generative techniques, demonstrating exemplary performance from generating coherent articles to poetry [9]. T5 views various NLP tasks as text generation challenges, simplifying the model's complexity while maintaining superior results [10].

Due to the outstanding performance of these models, many industries have started to recognize their potential value. From internet searches and ad recommendations to medical diagnoses and financial analysis, techniques based on the Transformer have become foundational in numerous contemporary applications.

Table 1: The statistics from the relation labels in subtask of Argument Relation Detection and Classification

	No detected (Relation=0)	support (Relation=1)	attack (Relation=2)
train	1600	3859	62
dev	200	482	8

3. METHODS

Identifying argumentative structures is a central focus of current research. The aim of this paper is to describe and demonstrate how advanced natural language processing techniques are employed to classify given sentences into 'Premise' and 'Claim'. To ensure the accuracy and efficiency of classification, we utilized the RoBERTa and BERT models from the HuggingFace's Transformers library for fine-tuning and employed GPT-3 turbo for sentence generation and auxiliary classification. More technical details are provided in the original papers of these models. In this section, we will delve into the models we chose, our training methodology, and how they are applied to the task of argument unit classification.

3.1 Imbalance Data Processing

Table 1 provides the descriptive statistics of label data from the subtask of NTCIR-17 FinArg-1 for Argument Relation Detection and Classification. According to Table 1, we found that there is an imbalance issue in the dataset. Figure 1 to 2 shows the visualization of descriptive statistics for the subtask of dataset, including the train and dev dataset. In order to solve the problem of imbalance data, we adopted the method of resampling the dataset for the experiment. Without changing the original content of the data, by using the oversampling method, the data of the "Attack" label can be

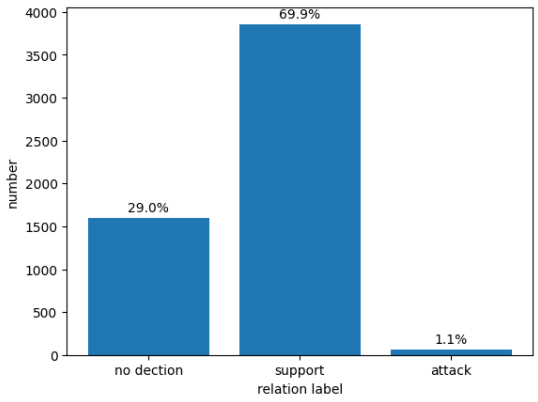

Figure 1: The relation distribution of label in train dataset from subtask of Argument Relation Detection and Classification

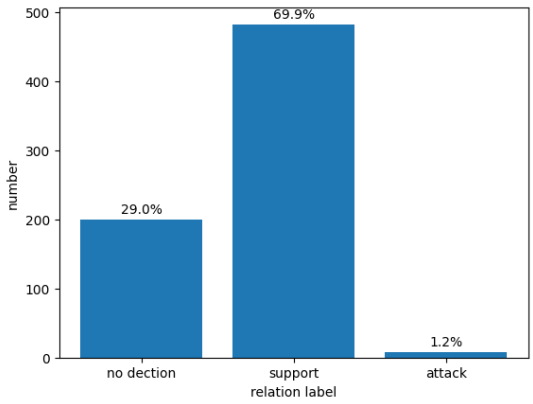

Figure 2: The relation distribution of label in dev dataset from subtask of Argument Relation Detection and Classification

resampling, thus making the distribution of the data more balanced, and allowing the model to better learn the classification of the three different labels. We have implemented two strategies for oversampling. The first involves augmenting the number of samples in "Attack" label to a level that is comparable to that of "No detected" label. This significantly mitigates the issue of data imbalance while concurrently minimizing the introduction of synthetic samples that could potentially lead to overfitting. As a result, the total number of samples for "attack" label approaches that of "No detected" label, thus ameliorating the problem of data imbalance. However, this does not completely balance the dataset since the number of samples in "Support" label remains predominant. Consequently, we adopted a second strategy, which entails increasing the number of samples in "attack" label to half of the comparable level of "No detected" label.

3.2 Selected Models

BERT (Bidirectional Encoder Representations from Transformers) was introduced by Google in 2018 and stands as a significant milestone in the field of natural language processing today [11]. It is built upon the Transformer architecture, which employs a self-attention mechanism, allowing the model to assign varying importance to words based on their context. A hallmark feature of BERT is its bidirectionality. Traditional language models, such as LSTM or GRU, learn text either from left to right or right to left. In contrast, BERT considers the context to the left and right of each word in the text simultaneously [12].

Its pre-training tasks, especially the MLM (Masked Language Model), require the model to predict words that are randomly masked in a sentence, ensuring that BERT grasps the overall structure and contextual meaning of the sentence. The NSP (Next Sentence Prediction) task trains the model to determine if two sentences are consecutive. Another innovative aspect of BERT is its input format. It can handle either a single sentence or a pair of sentences simultaneously, making it particularly effective for tasks like question-answering or textual entailment. Special tokens, such as [CLS] and [SEP], assist the model in identifying the start and end of sentences and differentiating between them [13].

RoBERTa (A Robustly Optimized BERT Pretraining Approach) was introduced by Facebook AI in 2019 and quickly garnered widespread attention in the field of natural language processing. While it builds on the original architecture of BERT, RoBERTa incorporates various enhancements and optimizations, achieving unprecedented results across multiple NLP tasks [14].

RoBERTa's training data significantly surpasses that of the original BERT, leveraging a larger dataset for pre-training. This allows the patterns. Additionally, RoBERTa's trainin model to capture a broader spectrum of linguistic nuances and g duration exceeds that of BERT, and it employs a larger batch size, both of which contribute to the model's deeper understanding of linguistic features. Structurally, RoBERTa deemed the NSP (Next Sentence Prediction) task as having limited impact on improving language model performance and thus opted to exclude it. However, the MLM (Masked Language Model) task was retained and further emphasized, given its prowess in aiding the model to comprehend semantic information within contexts [15]. Moreover, RoBERTa adjusted some of BERT's hyperparameters to further enhance the model's performance.

In several standard NLP benchmarks, such as GLUE, SQuAD, and SWAG, RoBERTa consistently showcased exceptional performance, breaking previous records multiple times [16]. This attests to its leading position in the domain of pre-trained language models. In summary, RoBERTa is not just an improved version of BERT but represents a new milestone in the natural language processing field, offering fresh perspectives and directions for subsequent research.

Albert (A Lite Bert for Self-supervised Learning of Language Representations) was also introduced by Google in 2019 to improve the efficiency of the previous BERT model by reducing model parameters [17]. Albert makes improvements in the previous design of BERT in three aspects: factorized embedding parameterization, cross-layer parameter sharing, and inter-sentence coherence loss. Albert gets significantly better results comparing to the BERT model.

FinBERT was proposed in 2019, which was a BERT-based large language model for natural language processing tasks in financial domain. In sentiment analysis tasks, FinBERT achieve state-of-the-art had a good performance with the evaluation on two financial sentiment analysis datasets [18].

The GPT (Generative Pre-trained Transformer) series, developed by OpenAI, is a set of advanced language models based on the Transformer architecture [19]. These models undergo unsupervised pre-training on vast amounts of text data, primarily aiming to predict the next word in a text. Once pre-trained, GPT models can be further fine-tuned to adapt to various natural language processing (NLP) tasks, such as text classification, sentiment analysis, and question-answering systems.

GPT-1, as the inception of the GPT (Generative Pre-trained Transformer) model series, marked the birth of pre-trained language models based on the Transformer architecture, with its primary applications in text generation and question-answering tasks. Its core objective is to utilize a plethora of unlabeled data to address the scarcity of annotated data in specific natural language understanding tasks.

GPT-2, in comparison to its predecessor, brought deeper innovations. The GPT-2 model, larger than GPT-1 with a parameter count reaching 150 million, can generate longer, more complex texts with increased accuracy. Use cases for the GPT-2 model include but are not limited to: automatic text summarization, machine translation, speech recognition, question-answering systems, chatbots, and more [20].

GPT-3 boasts 175 billion parameters and has achieved unprecedented results across various NLP tasks. Moreover, GPT-3 has been demonstrated to perform "few-shot learning," meaning it can learn new tasks by seeing only a few examples [21]. This makes

GPT-3 not only suitable for text generation but also for advanced applications like chatbots, automatic programming assistants, and knowledge question-answering.

GPT-3.5 Turbo, introduced by OpenAI, is an advanced language model based on the GPT (Generative Pre-trained Transformer) architecture [22]. It can accurately understand and generate natural language text. It's a more streamlined and high-performance variant of GPT-3 and is the model used by ChatGPT. Now, developers can fine-tune GPT-3.5 Turbo to produce better-customized models for their use cases and run these models at scale.

3.3 Fine-tuning Techniques

In the field of Natural Language Processing (NLP), deep learning models such as BERT, ALBERT, and their advanced version RoBERTa have demonstrated outstanding performance. This study employed these models for sentence classification tasks. Particularly for RoBERTa, with more data and longer training duration, it not only excels in various NLP tasks but also, through fine-tuning the original BERT training strategy, eliminates some of the limitations of previous versions. On the implementation level, we used RobertaTokenizer for text tokenization and RobertaForSequenceClassification for training and evaluating the model.

In the experimental design of our research, we utilized a 5-fold cross validation technique to assess and fine-tune the model's performance. This method divides the dataset into five subsets, with one serving as the validation set and the remaining four as training sets. The process is repeated five times, each time selecting a different subset for validation. This approach ensures a more stable and reliable performance estimate, ultimately confirming the stability of the model after fine-tuning. Additionally, we carefully adjusted several hyperparameters, including maximum sentence length, batch size, and training epochs, to enhance the model's learning efficiency. At the same time, to ensure a fair comparison with RoBERTa, we also used the same hyperparameter settings when fine-tuning the BERT model.

The model of Albert and Roberta-large had good performance in identify Identifying Attack and Support Argumentative Relations in Social Media Discussion Threads. In practical execution, we used AutoTokenizer for text tokenization and Automodel to train.

In addition, we also using the fine-tuning techniques on FinBERT. In the field of Natural Language Processing tailored for financial contexts, FinBERT exhibits exceptional efficacy. Its capabilities are further enhanced through the process of fine-tuning, rendering it even more effective for specialized tasks such as sentiment analysis and classification within the financial domain. During the fine-tuning process, we utilized AutoTokenizer from the Transformer library and specifically selected ProsusAI/finbert as the tokenizer. Concurrently, we employed AutoModelForSequenceClassification, also from the Transformer library, and opted for ProsusAI/finbert as the pre-trained model to be loaded.

3.4 GPT Generation Strategies and Optimization

Combining deep learning techniques with Natural Language Processing (NLP), OpenAI's ChatGPT API offers a powerful framework for conducting intricate text analysis. In this study, this technology has been specifically fine-tuned and applied for meticulous classification of sentences, particularly categorizing them as "claim" and "premise". Such classification not only assists researchers in understanding the key points of an argument but also reveals the foundational reasons supporting these main viewpoints.

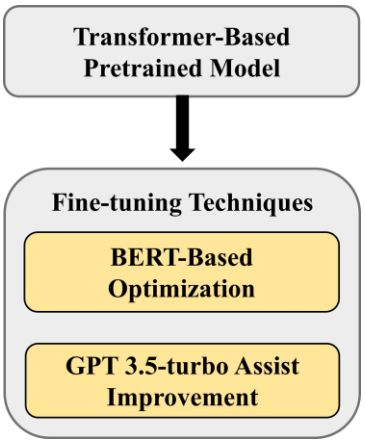

Figure 3: Proposed research architecture of IMNTPU at NTCIR-17 FinArg-1

In practice, this technique requires multiple steps to be fully implemented. By clearly defining sentiment labels, we provide the model with a clear classification target. These two labels - "claim" and "premise" - are crucial in this context, as they respectively represent the main points of an argument and the reasons supporting these points.

However, this is more than just a simple classification task. Properly identifying and classifying these elements helps to uncover the deep structure of the text, laying a solid foundation for subsequent analysis. For instance, distinguishing between the primary argument and its supporting reasons is vital for understanding and evaluating the quality of debates, the structure of academic papers, or other forms of professional literature [23].

Furthermore, the scalability and adaptability of this method make it highly practical in various scenarios. It can be applied not only to basic sentiment analysis but can also be expanded further to handle more classification labels or to undertake more in-depth text analysis [24].

3.5 Research framework

In this study, we introduced a comprehensive research framework specifically designed to address the Argument Unit Classification task in NTCIR-17 FinArg-1. This framework integrates three major strategies: fine-tuning techniques, GPT methodologies and applications, and model integration and evaluation. Initially, we adopted the BERT and RoBERTa models and fine-tuned them, leveraging their pre-training capabilities on large-scale corpora to capture subtle nuances and patterns within argumentative texts.

Subsequently, we further explored the potential application of GPT 3.5 Turbo, a model specifically designed for generative tasks. It was employed to generate classification labels for arguments, and its performance was assessed in comparison to traditional classification models for this task. Lastly, we undertook the integration and evaluation of the models, which included comparing the performance of various models on the test set and conducting an in-depth analysis of the advantages and disadvantages of each method. In summary, our research framework offers a holistic and robust approach to tackle the Argument Unit Classification challenge, laying a solid foundation for future research in this domain.

In our research concerning the FinArg-1 dataset, specifically for the task of Argument Relation Detection and Classification, we implemented a strategy of optimization through the application of

BERT-based algorithms. A comprehensive analysis and observation of the FinArg-1 data revealed that the task inherently necessitates multi-class sentence-pair classification. At first, to establish a robust foundation for our experiment, we initially chose a diverse set of BERT-based models to serve as our baseline. Secondly, we adopted a method of transfer learning, which was fine tuning. This fine-tuning was instrumental in tailoring the models to better suit the nuances of our specific research objectives. After the fine-tuning phase, we conducted a comparative assessment of all model variants using a carefully selected array of evaluation metrics to compare their relative performance and accuracy.

In Identifying Attacks and supporting augmentative relations within social media discussion threads in the NTCIR-17 FinArg-1 task, we employed a comprehensive research approach centered around the utilization of Albert and RoBERTa-large models, fine-tuning them to address the challenges.

4. EXPERIMENTS

In this study, our team delved deeply into the issue of argument unit classification. NTCIR-17 FinArg-1 provided three datasets: "Training Set", "Development Set", and "Test Set" [25]. We primarily used the "Test Set" to evaluate the performance of three IMNTPU models, with detailed results presented in Table 3, 4, 5. Among them, Micro-F1, Macro-F1, and Weight-F1 were our main evaluation metrics. Through these metrics, we could assess the model's performance across different argument unit categories. Furthermore, we explored various deep learning strategies, including the fine-tuning and application of BERT, RoBERTa, GPT 3.5 Turbo, FinBERT, AlBERT models.

4.1 Evaluation metrics

We focus on the problem of argument unit classification, aiming to precisely categorize a given sentence as either a "premise" or a "claim". For model evaluation, we place special emphasis on the three evaluation metrics recommended by the competition organizers: Micro-F1, Macro-F1, and Weight-F1. Micro-F1 considers the true positives, false positives, and false negatives of all categories, calculating the overall precision and recall. Macro-F1 is the average F1 score of each category, without considering the quantity of categories. Weight-F1 calculates the F1 score after considering the quantity or importance of each category. In addition to these three primary metrics, we also use Accuracy, Precision, F1-Score, and Recall as auxiliary evaluation metrics. These metrics not only comprehensively assess the overall performance of the model but also ensure that the model performs excellently across different argument unit categories.

4.2 Hyperparameter settings

In the training process of deep learning models, choosing appropriate hyperparameters, such as learning rate, max length, batch size, and number of epochs, is indispensable. These hyperparameters not only directly influence the speed of model training but are also closely related to the model's final performance and generalization capability. For instance, the learning rate determines the rate of model weight updates. The batch size affects memory usage and the model's convergence. The number of epochs pertains to how many times the model trains over the entire dataset. Proper hyperparameter configurations can allow the model to converge to an optimal solution more quickly, while avoiding issues of overfitting or underfitting. Table 2 details the hyperparameter settings chosen in our experiment, providing further reference.

Table 2 Hyperparameter Settings

Hyperparameter	Value
Learning rate	1e-5, 5e-5
Max Length	128, 256
Batch size	8, 16
Epochs	3, 4, 5

Table 3 Results of IMNTPU at NTCIR-17 FinArg-1 Argument Unit Classification

NTCIR-17 FinArg-1 Argument Unit Classification Development set				
Model	Micro-F1	Macro-F1	Weight-F1	Accuracy
IMNTPU-1 (BERT-base)	75.44%	75.31%	75.40%	74.82%
IMNTPU-2 (RoBERTa-base)	76.06%	76.05%	76.07%	75.64%
IMNTPU-3 (GPT 3.5-turbo)	56.97%	56.82%	56.70%	55.08%

Table 4 Results of IMNTPU at NTCIR-17 FinArg-1 Argument Relation Detection and Classification

NTCIR-17 FinArg-1 Argument Unit Classification Development set				
Model	Micro-F1	Macro-F1	Weight-F1	Accuracy
IMNTPU-1 (RoBERTa-base)	78.99%	47.36%	76.54%	78.55%
IMNTPU-2 (FinBERT)	82.61%	52.97%	82.14%	79.13%
IMNTPU-3 (BERT-uncased)	80.72%	50.73%	79.67%	78.55%

Table 5 Results of IMNTPU at NTCIR-17 FinArg-1 Identifying Attack and Support Argumentative Relations in Social Media Discussion Threads

NTCIR-17 FinArg-1 Argument relation in social media Development set			
Model	Micro-F1	Macro-F1	Wight-F1
IMNTPU-1 (Fintuned-Albert)	52.88%	34.77%	48.73%
IMNTPU-2 (Roberta-large)	48.71%	24.64%	40.50%

4.3 IMNTPU NTCIR-17 FinArg-1 Results

In the evaluation of the NTCIR-17 FinArg-1 Argument Unit Classification development set, we systematically compared three IMNTPU models and presented the detailed results in Table 3. The IMNTPU-1 model, utilizing the BERT-base architecture, achieved a Micro-F1 score of 75.44% and an accuracy of 74.82%. In contrast, the IMNTPU-2 model, based on the RoBERTa-base architecture, not only reached a Micro-F1 score of 76.06% but also improved its accuracy to 75.64%. This data clearly highlights the superiority and consistency of the RoBERTa architecture for this task. However, the IMNTPU-3 model, built on the GPT 3.5-turbo architecture, only managed a Micro-F1 score of 56.97%, showing a significant difference from the first two models. This result suggests that, for the specific task of Argument Unit Classification, the GPT 3.5-turbo architecture might not possess the efficiency of the BERT or RoBERTa architectures.

In Argument Relation Detection and Classification subtask, based on the performance of all the models, we picked the top-3 models and organized their experimental results into Table 4. Firstly, in our experiment, IMNTPU-1 was a fine-tune version of RoBERTa. It got 78.99% for Micro-F1, 47.36% for Macro-F1, 76.54% for Weight F1, and 78.55% for the accuracy. For IMNTPU-2, it was a fine-tuning edition of FinBERT, and it obtained 82.61% for Micro-F1, 52.97% for Macro-F1, 82.14% for Weight F1, and 79.13% for accuracy. At last, IMNTPU-3 was a fine-tuning form of BERT, it attained 80.72% for Micro-F1, 50.73% for Macro-F1, 79.67% for Weight F1, and 78.55% for the accuracy. Through all the experiment, we found that IMNTPU-2 achieved the best performance. In addition, FinBERT performed exceptionally well and is considered the most suitable model in the subtask of Argument Relation Detection and Classification.

The results of study show in Table 5. The IMNTPU-1 model with Roberta-large in three scores are all surpass the IMNTPU-2 model with the fintuned-Albert model. Therefore, we can find that RoBERta performs better in Identifying Attack and Support Argumentative Relations in Social Media Discussion Threads.

5. CONCLUSIONS

In this paper, we adopted three leading deep learning strategies of the current era. Initially, we fine-tuned the BERT and RoBERTa models, leveraging their pre-training capabilities on large-scale corpora to capture subtle differences and patterns within conversational texts. The powerful representational learning abilities of these transformer models enabled us to delve deep into the semantic structures of sentences, facilitating more precise classification.

Furthermore, we explored the potential of GPT 3.5 Turbo, an advanced language model designed specifically for generative tasks. By utilizing GPT 3.5 Turbo, we not only generated classification labels for arguments but also sought to understand how its performance in this specific task compares to traditional classification models.

The primary contribution of our research lies in offering a comprehensive solution to address the Argument Unit Classification challenge, coupled with a detailed analysis of the advantages and disadvantages of various methods. We not only employed mainstream fine-tuning techniques for classification but also ventured into using the GPT series models for binary classification. This diversified research approach allowed us to holistically evaluate the efficacy of various techniques in this specific task. Our findings will serve as a valuable reference for future research in this domain and lay a solid foundation for the further optimization of sentiment analysis techniques.

In multi-class sentence-pair classification, we fine-tuned BERT, FinBERT, and RoBERTa for the task. Through our experiments, FinBERT demonstrated the best performance on the FinArg-1 dataset. Over the course of our in-depth research, we developed a comprehensive understanding of the architectural nuances involved in multi-class classification tasks, specifically within the domain of sentiment analysis. By methodically analyzing inter-sentential relationships, we were able to uncover hidden semantic layers and nuances concealed within the textual data. This nuanced understanding significantly contributed to improving the efficiency and effectiveness of data analysis in the financial domain.

At the task of Identifying Attack and Support Argumentative Relations in Social Media Discussion Threads, we fine-tuned the Roberta and Albert model to capture subtle difference in two sentences and give them classification for ‘positive’, ‘negative’, and ‘neutral’. At the result, we found that RoBERTa outperformed other models in the task of discerning the relationship between two sentences.

6. ACKNOWLEDGEMENT

This research was supported in part by the National Science and Technology Council (NSTC), Taiwan, under grants MOST 110-2410-H-305-013-MY2, NSTC 112- 2425-H-305-002-, and NSTC 112-2627-M-038-001-, and National Taipei University (NTPU), Taiwan under grants 112-NTPU-ORDA-F-003, 112- NTPU-ORDA-F-004, USTP-NTPU-TMU-112-01, NTPU-112A413E01, and NTPU-112A513E01.

7. REFERENCES

[1] T. Loughran and B. McDonald, "Textual analysis in accounting and finance: A survey," *Journal of Accounting Research,* vol. 54, no. 4, pp. 1187-1230, 2016.

[2] N. V. Babu and E. G. M. Kanaga, "Sentiment analysis in social media data for depression detection using artificial intelligence: a review," *SN Computer Science,* vol. 3, pp. 1-20, 2022.

[3] A. Alsaeedi and M. Z. Khan, "A study on sentiment analysis techniques of Twitter data," *International Journal of Advanced Computer Science and Applications,* vol. 10, no. 2, pp. 361-374, 2019.

[4] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," *Decision Analytics Journal,* vol. 3, p. 100073, 2022.

[5] J. Li, L. Yang, B. Smyth, and R. Dong, "Maec: A multimodal aligned earnings conference call dataset for financial risk prediction," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 3063-3070.

[6] D. Matsumoto, M. Pronk, and E. Roelofsen, "What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions," *The Accounting Review,* vol. 86, no. 4, pp. 1383-1414, 2011.

[7] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.

[8] S. Alaparthi and M. Mishra, "BERT: A sentiment analysis odyssey," *Journal of Marketing Analytics,* vol. 9, no. 2, pp. 118-126, 2021.

[9] R. Dale, "GPT-3: What’s it good for?," *Natural Language Engineering,* vol. 27, no. 1, pp. 113-118, 2021.

[10] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research,* vol. 21, no. 1, pp. 5485-5551, 2020.

[11] M. G. Sousa, K. Sakiyama, L. de Souza Rodrigues, P. H. Moraes, E. R. Fernandes, and E. T. Matsubara, "BERT for stock market sentiment analysis," in *2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI)*, 2019: IEEE, pp. 1597-1601.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[13] M. D. Deepa, "Bidirectional encoder representations from transformers (BERT) language model for sentiment analysis task," *Turkish Journal of Computer and Mathematics Education (TURCOMAT),* vol. 12, no. 7, pp. 1708-1721, 2021.

[14] L. Zhao, L. Li, X. Zheng, and J. Zhang, "A BERT based sentiment analysis and key entity detection approach for online financial texts," in *2021 IEEE 24th International*

Conference on Computer Supported Cooperative Work in Design (CSCWD), 2021: IEEE, pp. 1233-1238.

[15] H. Tian *et al.*, "SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis," *arXiv preprint arXiv:2005.05635*, 2020.

[16] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[18] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv:1908.10063*, 2019.

[19] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[20] I. Beaver, "The success of Conversational AI and the AI evaluation challenge it reveals," *AI Magazine*, vol. 43, no. 1, pp. 139-141, 2022.

[21] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877-1901, 2020.

[22] J. Ye *et al.*, "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," *arXiv preprint arXiv:2303.10420*, 2023.

[23] C. Hidey, E. Musi, A. Hwang, S. Muresan, and K. McKeown, "Analyzing the semantic types of claims and premises in an online persuasive forum," in *Proceedings of the 4th Workshop on Argument Mining*, 2017, pp. 11-21.

[24] K. Kheiri and H. Karimi, "SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning," *arXiv preprint arXiv:2307.10234*, 2023.

[25] C.-C. a. L. Chen, Chin-Yi and Chiu, Chr-Jr and Huang, Hen-Hsen and Alhamzeh, Alaa and Huang, Yu-Lieh and Takamura, Hiroya and Chen, Hsin-Hsi, "Overview of the NTCIR-17 FinArg-1 Task: Fine-Grained Argument Understanding in Financial Analysis," in *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies*, Tokyo, Japan, 2023.