The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

Fine-Grained Argument Understanding with BERT Ensemble Techniques: A Deep Dive into Financial Sentiment Analysis

Eugene Sy, Tzu-Cheng Peng, Shih-Hsuan Huang, Hen-You Lin, and Yung-Chun Chang

Graduate Institute of Data Science, Taipei Medical University, Taiwan

{m946111012, m946111003, m946110008, m946111008, changyc}@tmu.edu.tw

Abstract

While argument mining has garnered attention over the years, its application in the financial sector remains nascent. This study presents a BERT-based ensemble learning approach tailored for sentiment analysis grounded in financial narratives, specifically focusing on unearthing arguments. For a nuanced analysis, we dissect the challenge into two pivotal subtasks using earnings conference call data: (1) Argument Unit Classification, and (2) Argument Relation Detection and Classification. Experimental results substantiate that our approach not only effectively forecasts both tasks but also outperforms the comparisons and achieve SOTA performance. This underscores the potential of our method in fine-grained argument understanding within financial analysis.

Keywords: Financial NLP, Ensemble Learning, Sentiment Analysis

1 Introduction

Financial technology (Fintech) has been developing for several years, and among them, Natural Language Processing (NLP) has gradually become a pivotal tool driving financial text analysis. Financial text analysis entails thorough examination of voluminous textual information within the financial domain, with the objective of unveiling the embedded emotions, sentiments, and logical frameworks, thereby offering supportive advice to investors and decision-makers. By taking advantage of natural language processing techniques, conducting sentiment analysis on financial texts effectively captures the emotional nuances within the content, subsequently predicting market trends and gaining insights into the oscillations of market participants' sentiments (Gupta, R., & Chen, M., 2020). Additionally, the application of natural language processing extends to natural language inference, dissecting the logical relationships among sentences in financial texts to uncover the structures of arguments and inferences, facilitating a profound comprehension of the underlying viewpoints and perspectives within the text (Chu et al, 2022). The amalgamation of these techniques not only enhances the precision and efficiency of financial text analysis but also provides an abounding source of information for decision-makers in the financial world, aiding them in making well-informed choices.

The textual data in the financial domain is vast and diverse, spanning various types such as analyst reports, earnings conference calls, and social media discussions. These texts not only encompass a wealth of information capable of influencing market sentiment and the formulation of investment strategies but also complicate and lengthen the process of effectively processing and analyzing this data. It is noteworthy that financial texts are replete with intricate technical terminology, posing a heightened level of challenge for natural language processing technology. Consequently, the specialized terminology within the financial domain presents hurdles for achieving the precision and efficacy of sentiment analysis and natural language inference, necessitating a more refined and intricate approach. Emotions within financial

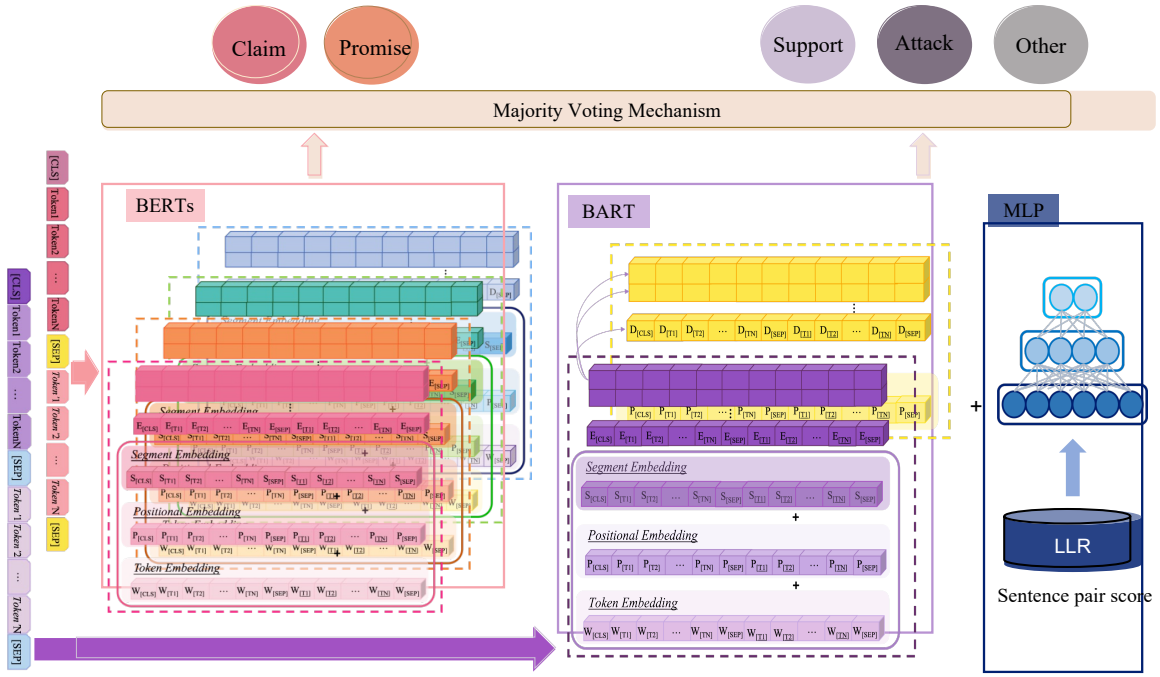

Figure 1. Overview of proposed method.

texts encompass a multitude of sentiments, ranging from positive and negative to neutral emotions, uncertainty, and emotional fluctuations. This wide spectrum of emotions significantly amplifies the intricacy of sentiment analysis. Moreover, an existing issue is the prevalence of uncertainty and ambiguity in financial texts. Market dynamics and financial events are frequently shaped by a blend of factors, leading to uncertainty and complexity that undermine the credibility and precision of sentiment analysis and inference. This concurrent uncertainty further intricately complicates financial projections and decision-making processes, underscoring the need for meticulous sentiment analysis to capture the nuances of emotional shifts in diverse contexts.

To overcome existing challenges, we employ two different approaches to sentiment analysis and to identify relationships between two different financial text datasets. First, we employ a method of categorizing texts as premises or claims in order to understand the point being conveyed. The second approach further focuses on extracting and evaluating the interrelationship between two sentences. Using NLP techniques and benefiting from pre-trained models, we aim to understand the interplay between language and context in these relationships. With this approach, we aim to be able to analyze and influence the choices of decision makers through the results of the proposed method.

This research enhances our understanding of text analysis within financial technology. Our pursuit is anchored on two paramount subtasks, elucidated as follows: (1) *Argument Unit Classification*: The primary objective of this task is to identify and categorize individual units or segments of arguments within the discourse found in earnings conference call data. This classification serves as a foundational step, enabling a granular breakdown of financial narratives. The precision in isolating these units paves the way for deeper comprehension and subsequent analysis. Recognizing the distinct units of arguments means that investors and stakeholders can better interpret the sentiments conveyed in these financial discussions. (2) *Argument Relation Detection and Classification*: This task aims to discern the intrinsic relationships between identified argument units. It's not merely about pinpointing the arguments but understanding the interplay between them. By classifying the nature and dynamics of these relationships, we gain insights into the coherence and flow of the financial narrative. Such an understanding is pivotal as it paints a clearer picture of the overall sentiment, aiding stakeholders in making informed decisions based on the interconnectedness of argumentative units.

Contributions of our paper can be summarized in the following:

- **Ensemble Technique Efficacy:** Our research showcases the effectiveness of employing an ensemble technique based on voting in the context of financial text analysis. Specifically tailored for argument unit identification and argument relation detection, this technique enhances the accuracy and reliability of our analysis. By harnessing the collective strength of multiple models, we offer a robust foundation for interpreting intricate financial discourse.
- **Adaptive Framework for Voting Mechanisms:** A noteworthy aspect of our work is the establishment of a flexible framework for implementing voting mechanisms across diverse language models. This adaptability empowers our methodology to be applied across various domains within the financial landscape. Our innovative approach reflects a commitment to versatility and extends the reach of our research.
- **Optimization Through Balancing Techniques:** An essential contribution lies in our utilization of targeted balancing techniques as part of data augmentation, optimizing language models before the voting process. This strategic refinement underscores our dedication to achieving superior outcomes. By employing these techniques in tandem with the ensemble approach, we demonstrate a rigorous methodology that enhances the reliability of our results.

2 Related Work

Sentiment analysis and opinion mining in financial texts has been significantly influenced by natural language processing techniques. Many research topics have explored the relationship between opinion mining and financial product prices in the financial domain, making numbers a crucial consideration in financial documentation. Therefore, a lot of research has been devoted to understanding the role of numbers in text analysis (Chen, C.C., 2019). In recent years, however, due to the rise of fintech, researchers' attention to text has exploded. This growing interest focuses on comprehensive analysis of investor sentiment to uncover more nuanced insights. While earlier studies have focused on coarse-grained analysis of market sentiment, often limited to binary bullish or bearish classifications, it is worth noting that the actual scope of financial market sentiment goes well beyond these binary labels (Chen, C.C., 2021).

The source of text data is primarily from discussions on social media platforms and discussions during earnings calls. In these datasets, the sentiment expressed is not just bullish or bearish. The focus is on the underlying relationships embedded in these discussions. In the context of natural language inference, well-known datasets such as MNLI (Williams et al., 2017) and SNLI (Bowman, S. R., 2015) stand out. These datasets aim to explore the implicit emotional connections between two sets of texts. Correspondingly, finarg-1 introduces datasets with parallel objectives, the key distinction lying in its task, which at its core is to determine the relationship between two sentences in a financial text. This approach is able to generalize the general sentiment of financial markets to specific themes.

With the advancements in natural language processing techniques such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020) have been introduced to financial text analysis, achieving notable outcomes in sentiment analysis. However, these generic models face challenges when handling domain-specific terms, specialized language, and logic within the financial domain. The FINBERT model developed by specifically targets financial sentiment analysis and effectively handles domain-specific terms and intricate logic present in financial texts (Araci, D., 2019). Additionally, some studies focus on analyzing financial social media data, which poses greater challenges for sentiment analysis due to the informal nature of social media texts (Chen et al., 2019). Moreover, in the context of financial market prediction, analysis of financial texts can be employed to apply machine learning or deep learning methods for predicting financial product prices (Leung et al., 2014) (Mehta et al., 2021).

By applying NLP techniques to the financial sector, researchers are afforded a deeper understanding of market sentiment, investor perspectives, and their impact on market trends and stock prices. Nonetheless, continued research

is imperative to develop more precise models that cater to the unique requirements of the financial domain, thereby achieving more accurate sentiment analysis and opinion mining. In this context, this paper has the potential to fill the existing knowledge gaps in the field, offering fresh avenues and insights for further in-depth research.

3 Method

The system architecture of our proposed method is shown in Figure 1. First, the Preprocessing Module is crucial in readying data for subsequent steps. Tailored to model needs and the voting process, it handles tasks like padding, truncation, and adding special tokens like [CLS] and [SEP] for Transformer-based models. This module also enhances the dataset by employing techniques like text augmentation, especially beneficial for limited datasets. The Voting Mechanism boosts the prediction performance for both tasks (Lin et al., 2022). We introduced two voting strategies: soft and hard voting. Soft voting considers the probability of each model's predictions, finalizing the most probable label. Hard voting, on the other hand, analyzes the majority prediction among models to arrive at a collective decision. We have placed the detailed code for both tasks on GitHub, allowing readers to gain a better understanding of the practical operational details.[1]

3.1 Argument Unit Classification (AUC)

Given an input argumentative sentence *s*, the objective is to develop a model *m* that accurately categorizes *s* into either the argument unit *A*={*claim*, *premise*} class. The challenge lies in ensuring that the model *m* possesses the ability to discern the nuanced differences between the two classes, optimizing for both precision and recall. The only preprocessing steps applied were those necessary for using a Transformer-based Language Model. These steps included text tokenization, adding [CLS] and [SEP] tokens, adjusting text length by padding or truncating (up to 512 tokens), and generating input IDs and attention masks for model training. In our pursuit of enhancing model performance, we further the adopt hard voting ensemble to combine various fine-tuned language models, including BERT, ROBERTA, ELECTRA, and FINBERT

3.2 Argument Relation Detection and Classification (ARC)

Given two sentences s_1 and s_2, the problem of this task can be defined as creating a method *m* is expected to determine the relationship between s_1 and s_2, categorizing it into one of the three relation class *R*={*support*, *attack*, *none*}, ensuring high accuracy in discerning the intricate inter-sentential relations. To begin, we paired the dataset, inherently composed of two discrete.
texts. This joint processing was facilitated through a transformer-based language model. A pivotal step here involved the integration of the [SEP] token, seamlessly interspersed between these paired texts, laying a robust groundwork for impending analyses.

In addition, there was an inherent imbalance in the dataset that was the most noticeable, which inevitably ushered in less-than-optimal results. To navigate this impediment, we undertook a multi-pronged strategy: (1) SMOTE Data Augmentation (RS): An experimental foray into the Synthetic Minority Over-sampling Technique (SMOTE) was undertaken. This technique artfully rebalanced the dataset by synthetically oversampling the minority classes. (2) Class Weighting (CL): Parallelly, we ventured into Class Weighting, applied judiciously to the loss function. The essence of this tactic was to allocate disparate weights to classes in alignment with their frequency. This inherently accorded higher significance to the more sparsely represented classes. In the integration of BERT for enhancing the performance, we employed a soft voting ensemble technique to amalgamate both BART and DEBERTA seamlessly.

Noteworthy, in the study by (Chang et al., 2022), there's compelling evidence illustrating the efficacy of the Log Likelihood Ratio (LLR) in generating and amalgamating linguistic patterns, leading to a substantial enhancement in predictive accuracy. Drawing inspiration from this revelation, we harness LLR to discern the significance of individual words nested within both sentences. This empowers us to cultivate distinguishing linguistic patterns rooted in their

[1] For the Augment Unit Classification, the code can be retrieved from: https://github.com/nlptmu/FinArg-1_AUC_FinSeq.
In addition, the relevant code of Argument Relation is available at: https://github.com/nlptmu/Finarg-1_ARC_-BDF4NLI

Table 1. The performance of compared methods for AUC task

	Precision / Recall / F_1-score (%)								
Methods	***Premise***			***Claim***			***Overall***		
BERT	77.86	75.32	76.57	73.77	76.42	75.07	73.77	76.42	75.82
RoBERTa	**79.53**	72.76	75.99	72.58	**79.38**	**75.83**	72.58	**79.38**	75.91
ALBERT	76.34	77.02	76.68	74.46	73.72	74.09	74.46	73.72	75.38
DistilBERT	76.82	75.65	76.23	73.64	74.88	74.25	73.64	74.88	75.24
FinBERT	78.41	73.92	76.10	73.00	77.60	75.23	73.00	77.60	75.66
ELECTRA	78.94	74.22	76.51	73.38	78.20	75.71	73.38	78.20	76.11
Our Method	77.95	**77.35**	**77.65**	**75.28**	75.92	75.60	**75.28**	75.92	**76.62**

Table 2. The performance evaluation for ARC task

	Precision/Recall/F_1-score (%)											
Methods	***Attack***			***Support***			***No-Relationship***			***Overall***		
BERT	0.00	0.00	0.00	70.00	**100**	82.32	**100**	0.05	0.01	56.65	33.50	27.79
FINBERT	0.00	0.00	0.00	70.59	99.59	82.61	80.00	4.00	7.60	50.20	34.55	30.08
DEBERTA	0.00	0.00	0.00	80.80	79.46	80.12	52.78	57.00	54.81	44.53	45.49	44.98
BART	0.00	0.00	0.00	82.93	91.70	**87.01**	73.89	58.00	**64.99**	52.27	49.90	50.69
Our Method	**100**	**12.50**	**22.22**	**84.18**	89.42	86.72	68.93	**61.00**	64.72	**84.37**	**54.30**	**57.89**

relationship scores. Thus Subsequently, this crafted feature space is seamlessly concatenated with the latent vector derived post an ensemble with BERT.

4 Experiments

4.1 Dataset and Setup

The datasets utilized in this study, derived from the NTCIR-17 FinArg-1 Shared Task (Chen et al., 2023), can be outlined as follows. These datasets are centered around textual content extracted from earnings conference calls within the financial domain, comprising two main categories. The first dataset is designated for the AUC task, which contains two labels "Claim" and "Premise" forms the basis of a binary classification task. It overall encompasses a total of 9,691 entries. This is subdivided into 4,613 Claim and 5,078 Premise. The second dataset is designed for the ARC task, which includes three possible relationships: "Attack," "Support," or "No Relationship." The dataset contains 8,148 entries, differentiated as follows: 4,596 are categorized as Support, 2,698 are classified as Attack, and 854 are labeled as having No Relationship.

In our experimental settings, we employed a 10-fold Stratified Cross Validation technique to ensure robustness in our model evaluation. Our primary metrics for evaluation include precision, recall, and the F_1-score, and to provide a holistic perspective of the model's performance across all classes, we used a macro-average approach. The hyperparameters were meticulously chosen based on preliminary testing and domain expertise. Specifically, the dropout was set at 0.35 to prevent overfitting, and we utilized the *AdamW* optimizer. For loss functions, the AUC task leveraged the *MSE* Loss, while the ARC task adopted the *CrossEntropy* Loss. The learning rate was set at 2e-05 for the AUC task and a conservative rate of 3e-07 for the ARC task. Furthermore, for epochs, the AUC task was trained for 2 epochs, whereas the ARC task extended to 30 epochs. These settings were strategically chosen to ensure optimal performance while minimizing potential overfitting, offering a comprehensive assessment of our model's capabilities.

4.2 Results of Argument Unit Classification

We conducted a comprehensive experiment to examine the performances of leading pre-trained

language models in the current landscape. Specifically, our comparative analysis encompassed models such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), DistilBERT (Sanh et al., 2019), FinBERT (Araci, D., 2019), and ELECTRA (Clark et al., 2020), shedding light on their relative efficacies and nuances in our designated tasks. This systematic evaluation serves as a foundation to discern the optimal model for specific applications in our study's context. As shown in Table 1, our method emerges as a standout performer with notably higher precision (77.95%) and recall (77.35%) values compared to the other models in Premise-related metrics. This disparity suggests that our approach not only captures a more substantial amount of pertinent information but also enhances accuracy in classifying premises. This capability significantly contributes to our method's impressive Premise Macro F_1-score of 77.65%.

Conversely, in the evaluation of Claim-related metrics, some of the other models, including Roberta, slightly outperform our method, showcasing elevated Precision (72.58%), Recall (79.38%), and consequently, superior Claim Macro F_1-scores. This implies that these models might possess a finer understanding of claim-related intricacies, leading to a more harmonious balance between precision and recall in this specific context. It would be prudent to investigate whether variances in training data or architectural nuances contribute to these disparities in claim-related performance.

Expanding the horizon to Overall metrics emphasizes the competitive landscape of all models involved. Our method maintains commendable Overall Precision (75.28%) and Overall Recall (75.92%), culminating in an Overall Macro F1-score of 76.62%. This high overall Macro F1-score is especially significant when considering real-world applications where robustness across multiple tasks is crucial. While our proposed method consistently delivers these high scores, there are trade-offs to consider, particularly in the area of Claim Macro F1-scores. The method is slightly outperformed by other models like RoBERTa in this metric. However, in real-world scenarios, a higher overall Macro F1-score is often more desirable than excelling in a single class. Therefore, a marginal decrease in Claim Macro F1-scores can be an acceptable trade-off for more balanced performance across various tasks and classes. The other models, including Roberta, also have their strengths, but our method's well-balanced blend of precision and recall indicates its consistent and versatile applicability.

The incorporation of the hard voting mechanism effectively consolidates their decisions and ensures comprehensive input analysis, resulting in significant improvements in precision, recall, and overall performance, as evidenced by our method's performance across the evaluation metrics.

4.3 Argument Relation Detection and Classification Experiments

In this experiment, building upon our previous assessment of BERT-based models, we further incorporated DEBERTA (He et al., 2020) and BART (Lewis et al., 2019) as baselines. This expansion in model comparison aims to underscore the benefits and efficacy of the methodology we propose, offering a more nuanced validation of our approach against a broader spectrum of contemporary language models. As shown in Table 2, our proposed method employed various pre-trained models for sentiment analysis of financial texts and conducted a detailed analysis of their performances. Of particular note, our method demonstrated outstanding performance in this task, with superior overall metrics compared to other methods, the performances of each model across different sentiment categories. BERT exhibited excellent precision but relatively lower recall and Macro F_1-score, possibly indicating the omission of certain actual sentiment instances. Although FINBERT achieved remarkable recall, the decrease in precision slightly affected its overall performance. DEBERTA, on the other hand, achieved a balance between precision and recall but relatively lagged behind other methods in overall performance.

The superior performance of our method is noteworthy across various aspects. Firstly, compared to the baseline model, our method showcased an impressive Precision, Recall, and Macro F_1-score in the Attack category, signifying its higher accuracy in predicting attack instances. Due to the severe class imbalance of the dataset, no baseline model was able to classify any attack instances, resulting in a zero metric score across

the board for this label. Furthermore, we also noted slightly lower recall and Macro F_1-score, suggesting potential missed attack samples. In the Support category, our method exhibited higher recall than other methods, indicating its ability to better capture actual Support instances. However, its comparatively lower precision implies some predictions might be misclassified as Support. Notably, in the category of No-Relationship, our method struck a good balance between precision and recall, excelling in this category. This implies its accurate identification of normal instances. Most importantly, our method's performance in overall metrics stands out. It achieved the best scores across all indicators, reflecting its balanced performance across multiple sentiment categories.

Taking into account all the factors, the remarkable performance of our method underscores its supremacy sentiment analysis applied to financial text. While alternative methodologies might exhibit strengths in specific areas, our approach consistently excels, showcasing its capacity to maintain a harmonized performance across a diverse spectrum of sentiment categories. This positions it as a potent and auspicious solution for sentiment analysis applications within the financial text. Furthermore, the success of our method can be attributed to the advanced capabilities of BART in grasping intricate contextual relationships within text. By leveraging the contextual understanding offered by BART, in combination with the integration of similarity LLR for sentence pairings, our method achieves a remarkable precision in accurately detecting instances of the 'attack' sentiment category. This nuanced approach allows our method to effectively address the challenges posed by the minority category, showcasing its ability to discern and classify even these relatively rare instances with a high degree of accuracy.

5 Conclusion Remarks

This research markedly advances text analysis in the financial technology domain, primarily focusing on two core subtasks: Argument Unit Classification and Argument Relation Detection and Classification. By meticulously segmenting and categorizing arguments in earnings conference call data and understanding their relationships, we offer stakeholders an enhanced clarity on financial sentiments. The contributions of this work include the demonstration of the ensemble technique's potency in financial text analysis, the introduction of a flexible framework for various voting mechanisms across language models, and strategic utilization of balancing techniques to optimize model performance. The experimental results demonstrate our proposed framework facilitates in-depth financial discourse, enabling stakeholders to make more informed decisions.

In future work, we aim to explore the integration of newer language models and expand our dataset to include diverse financial discourses from various global markets. Additionally, refining our balancing techniques and investigating real-time applications for instantaneous stakeholder insights are on the horizon. Our focus remains on enhancing accuracy and broadening the applicability of our methodology.

References

Gupta, R., & Chen, M. (2020, August). Sentiment analysis for stock price prediction. In *2020 IEEE conference on multimedia information processing and retrieval (MIPR)* (pp. 213-218). IEEE.

So, R., Chu, C. F. C., & Lee, C. W. J. (2022, May). Extract Aspect-based Financial Opinion Using Natural Language Inference. In Proceedings of the 2022 International Conference on E-business and Mobile Commerce (pp. 83-87).

Chen, C. C., Huang, H. H., Takamura, H., & Chen, H. H. (2019). Overview of the ntcir-14 finnum task: Fine-grained numeral understanding in financial social media data. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies* (pp. 19-27).

Chen, C. C., Huang, H. H., & Chen, H. H. (2021). From opinion mining to financial argument mining (p. 95). Springer Nature.

Williams, A., Nangia, N., & Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

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

Leung, C. K. S., MacKinnon, R. K., & Wang, Y. (2014, July). A machine learning approach for stock price prediction. In *Proceedings of the 18th International Database Engineering & Applications Symposium* (pp. 274-277).

Mehta, Y., Malhar, A., & Shankarmani, R. (2021, May). Stock price prediction using machine learning and sentiment analysis. In *2021 2nd International Conference for Emerging Technology (INCET)* (pp. 1-4). IEEE.

Sheng-Jie Lin, Wen-Chao Yeh, Yu-Wen Chiu, Yung-Chun Chang*, Min-Huei Hsu, Yi-Shin Chen, and Wen-Lian Hsu, "A BERT-based Ensemble Learning Approach for the BioCreative VII Challenges: Full-text Chemical Identification and Multi-label Classification in PubMed Articles," Database - The Journal of Biological Databases and Curation, 2022. (IF: 4.462, JCR: Q1)

Yung-Chun Chang, Chih-Hao Ku* and, Duy-Duc Le Nguyen, "Predicting aspect-based sentiment using deep learning and information visualization: The impact of COVID-19 on the airline industry," Information and Management, 2022.

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

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*

Chen, Chung-Chi and Lin, Chin-Yi and Chiu, Chr-Jr and Huang, Hen-Hsen and Alhamzeh, Alaa and Huang, Yu-Lieh and Takamura, Hiroya and Chen, Hsin-Hsi. (2023). Overview of the NTCIR-17 FinArg-1 Task: Fine-Grained Argument Understanding in Financial Analysis. Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan