Sentiment Analysis of News Articles on Gold in Vietnam

Nguyễn Ngô Thành Đạt, Nguyễn Phước Thắng, Nguyễn Anh Phi, Nguyễn Khánh Duy, Trần Quốc Khánh
Faculty of Information Science and Engineering, University of Information Technology,
Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{21522923, 21522590, 19522005, 21522003}gm.uit.edu.vn
khanhtq@uit.edu.vn

Abstract

Throughout human development, gold has been a longstanding medium of value exchange, and articles analyzing the gold price market have become increasingly abundant. However, there are still very few datasets about gold in the current research field, especially in Vietnam. In this paper, we introduce the "Sentiment Analysis of News Articles on Gold in Vietnam." This dataset is created to classify news articles about gold in Vietnam. It follows specific and easy-to-understand guidelines, with the label INCREASE representing news containing information indicating a rise in gold prices, and DECREASE representing news containing information about a drop in gold prices. The dataset was experimented on using three powerful pre-trained neural network models for Vietnamese: VisoBERT, CafeBERT, and PhoBERTv2. The best results were achieved with PhoBERTv2, with an Accuracy score is 0.9459 and an F1 score is 0.9474. Our dataset is available for research purposes. Please see more at the GitHub link: https://github.com/khanhsduyy/Crawl$_{Data_{Gold_{S}JC}}$.

1 Introduction

Sentiment analysis, also known as opinion mining, is a prominent application of Natural Language Processing (NLP) and machine learning. It involves the process of determining the emotional tone or subjective information from text data. This technique is widely used to understand the sentiment expressed in reviews, social media posts, news articles, and other forms of text data, providing valuable insights into public opinion and market trends

News articles on gold in Vietnam offer valuable market insights but are frequently affected by editorial biases, sensationalism, and inaccuracies. They can suffer from reporting delays, outdated information, and a tendency to oversimplify complex economic issues, focusing mainly on short-term events rather than long-term trends. The sheer volume of articles can lead to information overload and conflicting reports, making it challenging to discern reliable information. Additionally, there is a risk of market manipulation through biased reporting, and articles often reflect local cultural perspectives, potentially lacking a global viewpoint. Therefore, investors and analysts must critically evaluate and corroborate information from multiple sources, while considering the broader context, to make informed decisions.

In this research, we present a study on analyzing the sentiment in news articles related in Vietnamese gold market to improve the efficiency of other vital tasks on this data. Firstly, we collected SJC gold data from reputable financial websites. Next, we annotated the data to Increase or Decrease based on their title and content. After that, starting experiments with pre-processed data and raw data using proposed model checkpoints, the data included article title and content. Finally, we compare the results and obtain the suitable technique. Our contributions in this paper is integrating PhoBERTv2, ViSoBERT, and CafeBERT in model sequence classification to enhance sentiment analysis of gold market news articles in Vietnam by leveraging their diverse linguistic capabilities.

2 Related Works

The task of text classification, particularly in the financial domain, has seen significant advancements with the advent of pre-trained language models. Traditional methods relied heavily on feature engineering and classical machine learning algorithms, such as Support Vector Machines (SVMs) and Naive Bayes classifiers (1; 2). These methods, while effective to an extent, required substantial manual effort in feature extraction and often struggled with capturing contextual nuances of the text.

The introduction of deep learning techniques marked a paradigm shift in text classification. Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks, provided mechanisms to better capture sequential dependencies in text data (3; 4). Convolutional Neural Networks (CNNs) also found applications in text classification, leveraging their ability to extract local features from text (5; 6).

With the development of the Transformer architecture, models like BERT (Bidirectional Encoder Representations from Transformers) revolutionized natural language processing tasks, including text classification (7). BERT's bidirectional approach allowed for a more comprehensive understanding of context compared to unidirectional models (8). Subsequent models such as RoBERTa (9), ALBERT (10), and XLNet (11) built upon BERT's architecture, introducing various improvements in training efficiency and performance.

In the context of Vietnamese text classification, several pre-trained models have been specifically developed to address the unique linguistic characteristics of the Vietnamese language. PhoBERT (12) and ViBERT (13) are notable examples that have shown promising results in various NLP tasks. These models leverage large-scale Vietnamese corpora to enhance their contextual understanding.

More recently, ViSoBERT and CaFeBERT have been introduced, further improving upon their predecessors by incorporating additional training data and optimized architectures for Vietnamese text (14; 15). These models have demonstrated superior performance in benchmarks, highlighting their effectiveness in handling the complexities of the Vietnamese language.

The application of these models in the financial domain, particularly for tasks such as news classification and sentiment analysis, has gained traction. Studies have shown that pre-trained language models can effectively capture the sentiment and intent behind financial news articles, aiding in tasks such as market prediction and automated trading strategies (16; 17).

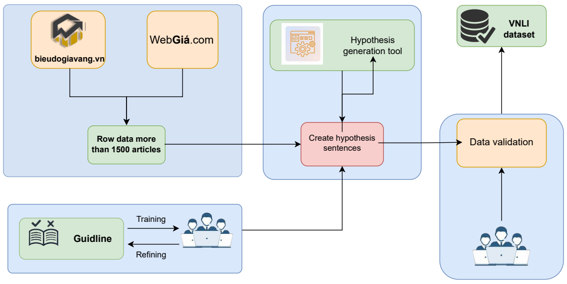
Overall, the continuous evolution of pre-trained language models and their adaptation to specific languages and domains underscores the importance of leveraging state-of-the-art techniques for improved text classification performance.

3 Dataset

3.1 Task definition

This dataset creation task is designed to compile comprehensive datasets encompassing two primary domains: SJC gold prices and blog content. The overarching objective is to facilitate in-depth analysis, shedding light on key aspects such as the fluctuation trends in gold prices and the sentiments expressed within blog posts.

3.2 Dataset creation


bieudogiavang.vn
WebGiá.com
Row data more than 1500 articles
Hypothesis generation tool
Create hypothesis sentences
VNLI dataset
Data validation
Guidline
Training
Refining

Hình 1: Figure 1: Dataset creation process

The dataset creation process consists of two primary steps:

3.2.1 Gold Price Collection

SJC gold prices are collected from the official webgia website. Selenium, in conjunction with BeautifulSoup, automates the data retrieval process to ensure accuracy and timeliness. The collected data includes essential details such as the SE number, time, bid price, and overall price of gold for each recorded instance.

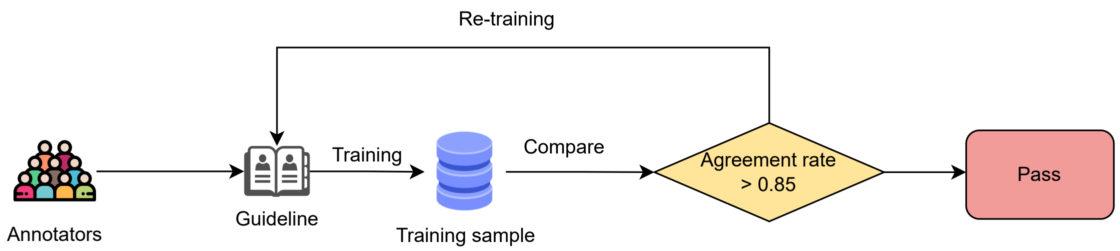
3.2.2 Blog Content Collection

Blog posts are sourced from the news section of bieudogiavang. Selenium and BeautifulSoup are employed to navigate through the website and extract the URLs of blog posts. Each URL is then visited to extract the title and content of the corresponding blog post. This meticulous process ensures the creation of a rich dataset encompassing a wide array of blog topics and their respective content for in-depth analysis.

These datasets serve as valuable resources for conducting various analytical tasks, including trend analysis of gold prices and sentiment analysis of blog posts. Additionally, the collected data can be leveraged for topic modeling and content categorization to gain deeper insights into market dynamics and audience interests.

3.3 Annotation Pipeline

In this annotation pipeline step, to increase consistency in labeling, the annotators are trained rigorously through the guidelines. In these guidelines,



Hình 2: The process of data annotation pipeline.

the annotators are clearly instructed on how to assign labels corresponding to each data point. The labels are as follows: INCREASE represents news containing information indicating an increase in gold prices, and conversely, DECREASE represents news containing information about a decrease in gold prices. You can see detailed examples in Table 1.

Next, each annotator will apply the knowledge gained from the guidelines to annotate 300 data points. After this, the agreement level will be evaluated. If the agreement level exceeds the predetermined threshold of 0.9, the annotators will proceed to label the next set of 300 data points. If the threshold is not met, the annotators will undergo retraining with the guidelines and reassess those data points until the entire dataset is labeled. Please see the detailed annotation pipeline process in Figure 2.

3.4 Data Quality Validation

Data validation is scientifically crucial in the sentiment analysis of news articles on gold to ensure accuracy and reliability, filter out irrelevant data, maintain consistency, and enhance model performance. This process leads to more precise market sentiment assessments, supporting better-informed investment decisions and strategies.

Key points:

- Data Collection Validation: We collected data from reputable online review platforms to ensure the reliability of our data sources as we collected data from webgia.com, bieudogiavang.com. The data include bid price, price, title, content, title and date of the articles. Ensured relevance by focusing on sources known for their coverage of financial markets, economy, and SJC gold-related news in Vietnam. Sampling articles from different time periods to capture temporal trends and fluctuations in sentiment

- Data Pre-processing Validation: In preprocessing stage, we conducted a sample check to ensure that text formatting inconsistencies, such as variations in capitalization and punctuation, were addressed uniformly across the dataset. Verified that any abbreviations or acronyms specific to gold-related terminology were appropriately expanded for clarity and consistency. Identified any missing or incomplete data fields, such as dates or price, through exploratory data analysis. Implemented strategies to address missing data, we use exclusion of incomplete records, while maintaining data integrity.

- Annotation Validation: In the dataset annotation phase, we use Label Studio, an open-source tool used for data labeling in artificial intelligence (AI) and machine learning (ML) projects. Using Label Studio to label the increase and decrease of gold prices based on the previous day's and the following day's gold prices can streamline the process of data annotation for gold price prediction models.

- Data Consistency Check: To maintain consistency in our dataset, we identified and removed duplicated objects using Pandas drop duplicates method. For format consistency, we standardized the format of dates and other relevant fields to ensure uniformity. Ensure that all articles are in a consistent language throughout the dataset.

In conclusion, the validation of the dataset for sentiment analysis of gold-related articles, incorporating both price data and article content, underscores its reliability and accuracy, as evidenced by strong agreement between annotators, high consistency in sentiment labeling, expert confirmation of accuracy, reliability of data sources, iterative refinement

Day	Title	Content	Lable
15/01/2024	Giá vàng hôm nay 15/1: Vàng sẽ tăng lại mức đỉnh vào tuần này?	Chuyên gia nhận định các yếu tố đang nghiêng về khả năng tăng và giá vàng có thể quay ...	INCREASE
02/11/2023	Giá vàng hôm nay 2/11: Đón tin xấu, giá vàng diễn biến ra sao?	Giá vàng giằng co trước những diễn biến khó đoán định của nền kinh tế Mỹ ...	DECREASE
02/12/2020	Giá vàng hôm nay ngày 02/12: Giá vàng đột ngột tăng trở lại	Giá vàng hôm nay ngày 02/12: Sau nhiều diễn biến rớt giá có thể coi là "thê thảm", giá vàng	INCREASE
20/04/2022	Giá vàng hôm nay 20/4: Hụt hơi, vàng quay đầu giảm sốc	Giá vàng năm châu không ngừng cắm đầu lao dốc trong phiên mua bán sáng hôm nay...	DECREASE

Bảng 1: Some value about features of the Gold news Dataset.

for quality assurance, alignment with external factors, and comprehensive documentation, thus affirming its suitability for robust sentiment analysis of news articles on gold in Vietnam.

3.5 Dataset Analysis

Introduction

In this section, we evaluate the dataset used in our study, which comprises 1555 records and 7 features. The dataset has been divided into training and test subsets in an 8:2 ratio. This evaluation covers the characteristics of the dataset, including feature analysis, summary statistics, and an assessment of missing values.

Data Overview The dataset consists of 1555 records and 7 features, encompassing numeric, categorical, and text variables. The features provide a comprehensive basis for our analysis and include:

- **TITLE**: Text feature representing the title of the record.
- **CONTENT**: Text feature representing the content.
- **BID_PRICE**: Numeric feature representing the bid price.
- **PRICE**: Numeric feature representing the price.

Feature Analysis and Summary Statistics

1. **Title Length**
   - **Average Length**: The average length of the titles is around 61 words.
   - **Range**: The longest title is 120 words, while the shortest is 31 words.
   - **Insight**: Titles vary significantly in length, which could impact text processing and analysis. The range indicates that some titles are much more detailed than others.
2. **Content Length**
   - **Average Length**: The average length of the content is 2012 words.
   - **Range**: The longest content is 4545 words, and the shortest is 121 words.
   - **Insight**: The content length also varies widely, which could affect text analysis and model training. Longer content pieces may provide more context and information.
3. **BID_PRICE**
   - **Maximum Value**: The highest bid price is 88.8.
   - **Average Value**: The average bid price is around 62.
   - **Missing Values**: There are 310 missing values for BID_PRICE.
   - **Insight**: The bid price shows a moderate range of values, and missing data needs to be addressed to prevent bias in analysis.
4. **PRICE**
   - **Maximum Value**: The highest price is 91.3.
   - **Average Value**: The average price is around 62.
   - **Missing Values**: There are 310 missing values for PRICE.
   - **Insight**: Similar to BID_PRICE, the price has a moderate range and significant missing data that must be handled.

Missing Data Analysis

- **BID_PRICE and PRICE**: Both features have 310 missing values, which is significant and represents approximately 20% of the dataset. Imputation or removal strategies must be considered to handle this missing data effectively.
- **TITLE and CONTENT**: These features have 629 missing values. Given the critical nature of these text features for analysis, addressing these missing values is crucial. The missing values constitute about 40% of the dataset, indicating a substantial amount of incomplete records.

Data Pre-processing Insights

- **Text Features**: The variation in lengths of TITLE and CONTENT suggests the need for pre-processing steps such as tokenization, stop-word removal, and potentially truncating or padding text to a uniform length for model input.
- **Numeric Features**: The presence of missing values in BID_PRICE and PRICE requires strategies such as mean/mode imputation or predictive imputation to ensure data integrity.

Challenges and Limitations

- **Missing Values**: A significant challenge is the high proportion of missing values, especially in TITLE and CONTENT. Strategies to impute or remove missing data will need careful consideration to avoid bias.
- **Imbalanced Lengths**: The wide range of lengths in text features may complicate text analysis and model performance, necessitating effective text pre-processing techniques.
- **Feature Diversity**: The dataset contains a mix of numeric, categorical, and text features, requiring a multifaceted approach to pre-processing and analysis.

Conclusion

The dataset presents a rich mix of features suitable for a variety of analyses, yet poses challenges such as significant missing data and varying lengths of text features. Addressing these issues through careful pre-processing and validation steps will be crucial to ensure the reliability and validity of subsequent analyses. Overall, the dataset holds potential for insightful analysis, provided that the identified challenges are adequately managed. This analysis highlights significant variability in the lengths of titles and content, as well as considerable missing data in both pricing and textual fields. Addressing the missing values will be crucial for accurate analysis and modeling. The statistics on lengths and prices provide a foundational understanding for further data processing and feature engineering.

4 Empirical Evaluation

4.1 Baseline Models

1. We present PhoBERTv2, trained exclusively on Vietnamese language data, exhibits high proficiency in understanding and processing Vietnamese text, allowing it to capture subtle nuances and context-specific information present in news articles about the Vietnamese gold market, consequently enhancing the accuracy and effectiveness of sentiment analysis on Vietnamese text.
2. ViSoBERT is trained on Vietnamese and English data, boasts multilingual capabilities, enabling effective text processing in both languages; its shared knowledge between the two languages during training enhances performance across diverse linguistic contexts, while its proficiency in capturing contextual information and semantic nuances enhances analytical capabilities, particularly in sentiment analysis tasks.
3. CafeBERT is trained on domain-specific data related to the culinary domain, possesses specialized knowledge in areas such as terminology and context, enhancing its analysis of news articles about the Vietnamese gold market, particularly in discussions of economic trends, market dynamics, and investment insights. Its expertise in the culinary domain offers unique perspectives that complement analyses by PhoBERTv2 and ViSoBERT, enriching the overall sentiment analysis process.

4.2 Settings

General settings: max_length = 256, lr = 10^{-5}, num_labels = 2, num_warmup_steps = 0,
seed = 42.

- PhoBERT: batch_size = 16, epochs = 5
- ViSoBERT: batch_size = 16, epochs = 5
- CaFeBERT: batch_size = 8, epochs = 5

4.3 Evaluation Metrics

- **Accuracy**: The ratio of correct predictions to the total number of predictions.
- **Precision**: The ratio of true positive predictions to the total number of positive predictions.
- **Recall**: The ratio of true positive predictions to the total number of actual positives.
- **F1-Score**: The harmonic mean of precision and recall, representing the balance between precision and recall.

4.4 Experimental Results

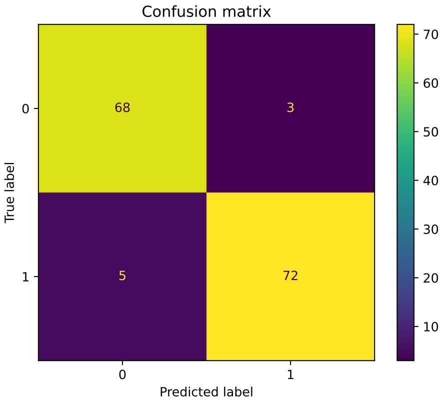
The results from Table [2] indicate several key observations:

- **High Performance of Pre-trained Models**: The pre-trained models, including PhoBERT, ViSoBERT, and CaFeBERT, all demonstrate high performance across various metrics. This highlights the effectiveness of leveraging pre-trained language models for the task of classifying news articles related to gold price movements.
- **Improvement with Preprocessed Data**: The use of preprocessed data significantly improves the performance of the models. This is evident from the increased accuracy and F1-score when comparing models trained on pre-processed data versus non-preprocessed data. The preprocessing steps likely help in reducing noise and enhancing the quality of the input data, leading to better model performance.
- **Best Performance by PhoBERT**: Among the tested models, PhoBERT shows the best performance with an accuracy of 94,59%, F1-score of 94,74% when trained on preprocessed data. This suggests that PhoBERT is particularly well-suited for understanding and processing Vietnamese text in the context of financial news, making it the most effective model for this specific task.

4.5 Result Analysis

The analysis results below are based on experiments that yielded the best results using PhoBERTv2 with input as Title + Content, conducted on a pre-processed and balanced dataset. For a detailed analysis, see the confusion matrix in Figure 3. The confusion matrix and derived metrics suggest that the model performs exceptionally well in classifying news related to gold. The high accuracy, precision, recall, and F1 score demonstrate that the model has a strong ability to correctly identify both positive and negative instances with minimal errors. This robust performance implies that the model is reliable and effective for this classification task.

However, there are still errors, especially with label decrease, where the model mostly misclassifies this label. This ambiguity stems from errors in data labeling, for example, on May 11, 2023, with the title "Giá vàng hôm nay 11/5 vọt lên cao rồi lao dốc"that mean “Gold price today 11/5 surged then plummeted,” which was labeled as Increase, but the purpose of the article was to indicate that the gold price was going down. Similarly, on May 14, 2024, "Giá vàng hôm nay 14/5 liên tục biến động vàng sjc bất ngờ giảm mạnh trước phiên đấu thầu"mean “Gold price today 14/5 constantly fluctuates, SJC gold unexpectedly drops sharply before the auction session.” The ambiguity in vocabulary leads to incorrect labeling and results that are not as expected.



Hình 3: Confution matrix

5 Conclusion and Future Work

5.1 Conclusion

In this study, we investigated the effectiveness of pre-trained language models, specifically PhoBERT, ViSoBERT, and CaFeBERT, for the task of classifying news articles about gold price movements into "Increase"and "Decrease"categories. Our empirical evaluation demonstrated that these models achieve high performance across multiple metrics, with PhoBERT achieving the best results. The use of preprocessed data significantly improved the models’ performance, underscoring the importance of data quality in natural language processing tasks. These findings suggest that leveraging pre-trained language models, along with thorough data preprocessing, can substantially enhance the accuracy and reliability of financial text classification.

5.2 Future Work

Future research could expand on this work in several directions:

		Not Pre-process				Pre-process			
		Not Balance		Balance		Not Balance		Balance	
Model	**Metrics**	**Title**	**Title+Content**	**Title**	**Title+Content**	**Title**	**Title+Content**	**Title**	**Title+Content**
VisoBERT	**Acc**	0.9038	0.8846	0.9054	0.8716	0.9038	0.8526	0.9122	0.8851
	Precision	0.8784	0.8354	0.9200	0.8372	0.8784	0.8077	0.9324	0.8750
	Recall	0.9155	0.9296	0.8961	0.9351	0.9155	0.8873	0.8961	0.9091
	F1 score	0.8966	0.8800	0.9079	0.8834	0.8966	0.8456	0.9139	0.8917
CafeBERT	**Acc**	0.9359	0.9232	0.9122	0.9189	0.9295	0.9231	0.9122	0.9324
	Precision	0.9178	0.9213	0.9444	0.9452	0.9054	0.9041	0.9324	0.9467
	Recall	0.9437	0.9315	0.8831	0.8961	0.9437	0.9296	0.8961	0.9221
	F1 score	0.9306	0.9212	0.9128	0.9200	0.9241	0.9167	0.9139	0.9342
PhoBERTv2	**Acc**	0.9231	0.9359	0.9257	0.9392	0.9359	0.9359	0.9257	**0.9459**
	Precision	0.9041	0.9178	0.9459	0.9474	0.9067	0.9178	0.9459	**0.9600**
	Recall	0.9296	0.9437	0.9091	0.9351	0.9577	0.9437	0.9091	**0.9351**
	F1 score	0.9167	0.9306	0.9272	0.9412	0.9315	0.9306	0.9272	**0.9474**

Bảng 2: Performance metrics of pre-trained models on various datasets.

- **Expanding the Dataset**: Collecting a larger and more diverse dataset of financial news articles could further improve the robustness and generalizability of the models.

- **Exploring Advanced Models**: Investigating the use of newer and more advanced models, such as GPT-3, T5, or other state-of-the-art language models, could potentially yield better performance.

- **Domain-Specific Fine-Tuning**: Fine-tuning the models on domain-specific corpora related to finance and economics might enhance their understanding and classification accuracy for specialized content.

- **Ensemble Techniques**: Applying ensemble methods to combine the strengths of multiple models could result in improved performance and robustness.

- **Enhanced Preprocessing Techniques**: Exploring more sophisticated preprocessing techniques, such as named entity recognition and sentiment analysis, could further refine the input data and boost model accuracy.

References

[1] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine Learning: ECML-98*, pp. 137–142, 1998.

[2] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752. Madison, WI, 1998, pp. 41–48.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[5] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.

[6] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," *arXiv preprint arXiv:1412.1058*, 2014.

[7] 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.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[10] 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.

[11] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, 2019.

[12] D. Q. Nguyen and T. V. Nguyen, "Phobert: Pre-trained language models for vietnamese," *arXiv preprint arXiv:2003.00744*, 2020.

[13] H. Nguyen, A. Le, and X.-S. Vu, "Vibert: A pre-trained language model for vietnamese text processing," *arXiv preprint arXiv:2003.06877*, 2020.

[14] D. Q. Nguyen and M. Le, “Visobert: A pre-trained language model for vietnamese social media text,” *arXiv preprint arXiv:2106.11007*, 2021.

[15] T.-H.-P. Tran and M.-T. Nguyen, “Cafebert: A pre-trained language model for vietnamese cafe reviews,” *arXiv preprint arXiv:2106.11009*, 2021.

[16] Z. Hu, W. Liu, J. Bian, X. Liu, H. Liu, and Z. Ren, “Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction,” *arXiv preprint arXiv:1811.06173*, 2018.

[17] Y. Yang, N. Chawla, and C. K. Reddy, “Finbert: A pretrained language model for financial communications,” *arXiv preprint arXiv:2002.10860*, 2020.