ISSN: 1859-3690

Journal

FINANCE - MARKETING RESEARCH

UNIVERSITY OF FINANCE – MARKETING

Vol. 15 Issue 2 - Year 2024

JOURNAL OF FINANCE – MARKETING RESEARCH
Vol. 15, Issue 2, March 2024

CONTENTS

Journal of Finance – Marketing Research; Vol. 15, Issue 2; 2024
ISSN: 1859-3690
DOI: https://doi.org/10.52932/jfm.vi2

Journal of Finance – Marketing Research
http://jfm.ufm.edu.vn

A SEMI-SUPERVISED APPROACH FOR VIETNAMESE STOCK NEWS CLASSIFICATION WITH DEEP LEARNING

Nguyen Minh Nhat[1*], Tran Huynh Minh Tan[1]
[1]Ho Chi Minh University of Banking (HUB), State Bank of Vietnam, Vietnam

ARTICLE INFO

DOI:
10.52932/jfm.vi2.500

Received:
February 12, 2024
Accepted:
March 20, 2024
Published:
March 25, 2024

Keywords:
BERT;
Deep learning;
Semi-learning;
Stock news;
Stock movements.

ABSTRACT

Stock-related news and articles on Vietnamese economic websites and blogs are rapidly increasing, but they are mixed with entertainment news, miscellaneous topics, and advertisements. This makes it annoy for real investors and analysts who only focus to find and analyze the stock-related information that matters (Boudoukh, 2013). This research introduces a novel method for automatically labeling the relevance of news articles to the stock market, based on a set of criteria derived from financial domain knowledge. In addition, this study also develops a deep learning classifier model that leverages the BERT architecture and the Vietnamese language model (*viBERT*) (Tran, 2020) to achieve high accuracy and efficiency in scoring the stock market news. This approach helps investors and analysts to filter out the irrelevant content on Vietnamese economic websites and access the most useful information for their mainstream analysis of stock movements.

*Corresponding author:
Email: nhatnm@hub.edu.vn

1. Introduction

Forecasting stock market trends has consistently held a certain appeal for researchers. While numerous research efforts have been undertaken to understand stock movements, no method has yet been found to predict stock price movements with complete accuracy (Khan, 2020).

Recent studies have shown the significant impact of news sentiment on stock prices and returns, this research successfully built trading strategies, which is out performance over other baseline methods (Allen, 2019). However, obtaining and filtering relevant news articles for stock-related sentiment analysis can be a challenging task. In order to achieve accurate sentiment analysis for stock news require the availability of precise and high-quality stock-related articles (Sun Y. M., 2018). Therefore, the primary goal of this research is to construct a stock news classifier that employs a cost-effective and low- cost labeling the dataset for training model. This approach seeks to overcome the challenges associated with the labor-intensive and costly process of obtaining labeled instances, which typically necessitates the involvement of skilled human annotators (Zhu, 2005).

According to one study (Gidofalvi, 2001), each article in a training set of news articles is categorized as 'up', 'down', or 'unchanged' based on the stock's movement during a time frame around the article's publication. For instance, if the stock price rises above a certain threshold within a specified period following the article's release, it is labeled as 'up'. Conversely, if the stock price falls below a certain threshold, the article is marked as 'down'. Articles are labeled as 'unchanged' if the stock price does not meet these criteria. The research utilized a naive Bayesian text classifier to determine the likely movement class for each article.

Another study (Villamil, 2023) employed a similar approach; however, instead of utilizing a fixed threshold, they determined the labels based on the average and standard deviation of stock price changes. In summary, this study applied embeddings and bidirectional recurrent neural networks to elucidate the dependencies between news articles and stock price movements, offering valuable insights for investors.

Sentiment analysis is a powerful tool for stock trading and investment, as it can capture the emotions and opinions of the market participants (Van de Kauter, 2015). However, sentiment analysis is only meaningful if it is based on relevant and high-quality articles that are related to stock only. Otherwise, the analysis might be biased or inaccurate, leading to poor investment decisions. Especially firm-specific news articles play a more significant role in influencing investor-trading activities compared to general news on social media (Qing Li, 2014).

This study (Duong, 2016) proposes a novel method for predicting the stock market trends in Vietnam using both stock news and stock prices of VN30 index. The experiment date was collected news articles from three major financial websites and represent them as feature vectors. The author applied a machine learning technique called stacked long short-term memory (LSTM) to integrate the news articles with the stock prices in our prediction model. The author claim that our method achieves high accuracy in VN30 index trend prediction. The research also compares our method with previous works that use either market news or historical stock prices for prediction. The result show that our method outperforms them in terms of accuracy and error rate. Their paper contributes to the literature on stock market prediction by using news articles as an important factor that influences investors'

behavior. Our paper also demonstrates the applicability of our method in the Vietnamese stock market, which has received little attention from previous research. The Vietnamese stock market has been on a strong upward trend since late 2023, thanks to the recovery of the domestic economy, the low interest rate environment, and the positive impact of supportive policies from the government and the central bank (Office, 2023). Foreign investors are expected to increase their participation in the Vietnamese stock market in 2024, as the market offers promising opportunities and benefits from the Fed's loose monetary policy (Tran Duc Anh, 2023). Hence, it is crucial to create a steady habit of producing high-quality articles that concentrate on stock only and revise them frequently to match the current market movements and occurrences. This continuous process ensures the reliability and precision of the analysis, which can facilitate more knowledgeable and assured investment choices.

2. Literature review

Financial news articles are believed to influence stock price returns. Prior research has often focused on analyzing the word patterns within news articles to discern the underlying relationship between these patterns and stock price fluctuations (Xiaodong Li, 2014). Indeed, within the Vietnamese stock market, the sentiment of the news—whether it is perceived positively or negatively—can exert a comparable and substantial effect on the volatility of stock prices. As observed in financial markets globally, optimistic news such as robust corporate earnings or positive economic indicators tend to drive stock prices upward. Conversely, adverse news, including disappointing earnings reports or economic recessions, generally result in stock price declines (Nguyen Van, 2015).

To effectively train traditional machine learning classifiers, labelled data is essential, pairing features with their corresponding labels. However, acquiring labelled instances often proves to be challenging, costly, and time-intensive, necessitating the expertise and diligence of skilled human annotators. Conversely, while unlabeled data can be collected with relative ease, it has historically been underutilized. Semi-supervised learning addresses this issue by leveraging a substantial volume of unlabeled data in conjunction with the available labelled data to construct classifiers that are both more reliable and accurate. A primary advantage of semi-supervised learning is its potential to reduce the human labor required for data labelling significantly, concurrently enhancing the precision of the models. This dual benefit has spurred considerable interest in semi-supervised learning, both from theoretical and practical standpoints, across diverse fields (Zhu, 2005; Devlin, 2018; Vaswani, 2017).

An alternative semi-supervised learning method, known as co-training, capitalizes on unlabeled data to enhance the efficacy of supervised learning algorithms. The core concept involves utilizing two disparate feature sets or views of the data, such as the textual content and hyperlinks on a web page. Two classifiers are independently trained on each view using a limited subset of labelled data. Subsequently, these classifiers are employed to confidently label the unlabeled data, and the newly labelled instances are incorporated into the training set of the opposing classifier. This iterative process continues until the pool of unlabeled data is exhausted or a predefined criterion is satisfied. Theoretical analysis of this method can be conducted within the PAC (Probably Approximately Correct) learning framework, which evaluates the accuracy and efficiency of learning algorithms. Empirical evidence also substantiates that co-training can significantly refine the classification of web pages when applied to real world data (Blum, 1998).

Forecasting stock market trends is an intricate endeavor that necessitates a deep comprehension of the interplay between news events, historical data analysis, and their collective impact on stock price dynamics. The volatile and unpredictable nature of stock prices further complicates this task. A notable gap exists in the literature regarding a holistic overview of the stock prediction landscape. Addressing this gap, an exhaustive survey has been conducted. This survey delineates all critical terms and phases inherent to the general stock prediction methodology, along with its inherent challenges. It encompasses a detailed literature review, spanning data pre-processing techniques, feature extraction, prediction methodologies, and the future trajectory of news-based stock prediction. The survey highlights the superiority of structured text features over unstructured and superficial counterparts and delves into the application of opinion extraction techniques. It also emphasizes the integration of domain expertise in both text feature extraction methods. Furthermore, the survey accentuates the pivotal role of deep neural network-based prediction models in unveiling the latent correlations between textual and numerical data. This survey is pivotal and groundbreaking as it formulates a comprehensive framework for stock market forecasting and critically assesses the merits and drawbacks of extant methodologies. It also unveils a plethora of open issues and research avenues, offering substantial contributions to the research community (Usmani, 2021).

Therefore, the creation of high-quality financial news data is crucial, as it can yield valuable insights into stock market trends. By scrutinizing news events and their influence on stock prices, investors can make well-informed decisions regarding the purchase and sale of stocks. Nonetheless, the pronounced volatility of stock prices poses a challenge to precise trend prediction. The survey referenced herein offers a comprehensive framework for stock market forecasting and delineates the strengths and limitations of current methodologies. It also introduces a broad spectrum of unresolved issues and research directions that could enhance the predictive accuracy of stock market trends. Our study contributes further to this analysis by demonstrating that adverse news disproportionately escalates volatility compared to positive news, and by affirming a positive correlation between returns and volatility in the Vietnamese market, particularly in terms of the impact on conditional volatility (Nguyen Van, 2015).

3. Methodology

Our research is centered on automating the pre-processing tasks for label generation, specifically within the context of authenticating news articles related to stock market classification. The goal is to meticulously prepare data labels and develop a model that effectively filters out generalized news, advertisements, and irrelevant content from disparate domains.

3.1. Ideas

The main idea of this research is to standardize and develop a labeling method for stock news data by isolating and scraping data exclusively from stock-related websites. Naturally, the retrieved data will predominantly pertain to stocks. However, there is still a lack of confidence to assert definitively that the obtained data is indeed relevant to the stock market.

Therefore, this research proposes a solution by employing pure keywords related to the stock market domain. These include stock codes, listed company names, specialized stock market terminology, and so far. The objective is to selectively choose high-quality stock-related news using a matching and rules-based method (Triguero, 2015).

For instance, by reviewing the Table I, which contains news data along with specific keywords that any professional investor would immediately recognize as highly relevant to the stock market. This serves as the research's discovery for labelling the dataset of stock-related news. The assumption is that news articles lacking these stock-related keywords are considered less connected to the stock market domain.

The advent of BERT (Bidirectional Encoder Representations from Transformers) (Wu, 2020), which has the capability to generate contextualized word vectors, marks a pivotal moment in the advancement of text classification and other Natural Language Processing (NLP) technologies. BERT's ability to consider the context of words within a sentence has revolutionized language representation models, leading to substantial improvements in various NLP applications by capturing nuanced semantic relationships and contextual information. Therefore, for future stock news classification, this research focuses on constructing a dataset with genuine quality and semantic emphasis on the stock market. The goal is to leverage the power of BERT by creating a dataset rich in stock-related content.

Table 1. Example of news related to stock market

Text	Label
Tháng 11/2023, Tập đoàn Hòa Phát (mã chứng khoán:HPG) đã sản xuất 623.000 tấn thép thô, tương đương tháng trước	relevant
Sản lượng thép của Hòa Phát trong tháng 12/2023 cao nhất 21 tháng, tiếp tục tăng giá bán	relevant
Ngành thép dự báo tăng trưởng 10% trong năm 2024	relevant
VIB và hành trình tiên phong về công nghệ thẻ tín dụng	irrelevant
Suzuki Jimny chốt lịch ra mắt tại Việt Nam dù đã có nhiều xe giao khách, đại lý báo giá từ 789 triệu, có bản độ sẵn cho khách thích G63	irrelevant
Tuần tăng điểm mạnh nhất của VN-Index trong 14 tháng	relevant

Note: *Symbol **The purpose of collected data is NOT to promote any mentioned companies/products. This data was provided by (STAG Vietnam, 2024)*

3.2. Experiment Implementation Pipeline

The implementation follows the steps outlined in the procedure depicted in flowchart Figure 1.

Data source crawler: Data can be obtained stock-specific news from financial news websites and news aggregators such as vnEconomy, Cafef, Vietstock, VietnamBiz, etc. Alternatively, you can use specialized financial news APIs to get real-time data on stocks and other financial instruments.

HTML Preprocessing: One common approach is to utilize a combination of libraries and tools for web scraping and HTML parsing.

The Boilerpipe library (Kohlschütter, 2022) was utilized to detect and extract the complete text from webpages. This Java library offers algorithms designed to identify and eliminate extraneous 'clutter'—such as boilerplate and templates—that surrounds the principal text of a webpage. The algorithm addresses the significant challenge of filtering out irrelevant

text from web articles, a critical concern for information retrieval systems (Francisco Viveros -Jiménez, 2018).

Stock Company and Symbol Recognition: To enhance the relevance of our article classification, we utilize keywords and phrases pertinent to the stock market, listed company entities, and the financial domain. This includes, but is not limited to, company names, stock ticker symbols, industry specific terms, and financial indicators such as 'HOSE,' 'HPG,' 'VNM,' and 'cổ phiếu.' These terms serve as filters to segregate articles into a strongly relevant group, ensuring that our data set is focused and pertinent to our research objectives.

The experiment incorporated a Named Entity Recognition (NER) approach to pinpoint and extract entities such as listed company names, stock symbols, and other financial identifiers from news articles. This ensures that the news articles containing these entities are directly related to the stock market or trading activities. The NER model was trained using a 'lazy' approach, which does not require labeled data, as detailed in the research by (Lison, 2020). The authors employed labeling functions to automatically annotate texts within the target domain, a concept that forms the cornerstone of our study. Main concept was descrised the flow chart in Figure 1.

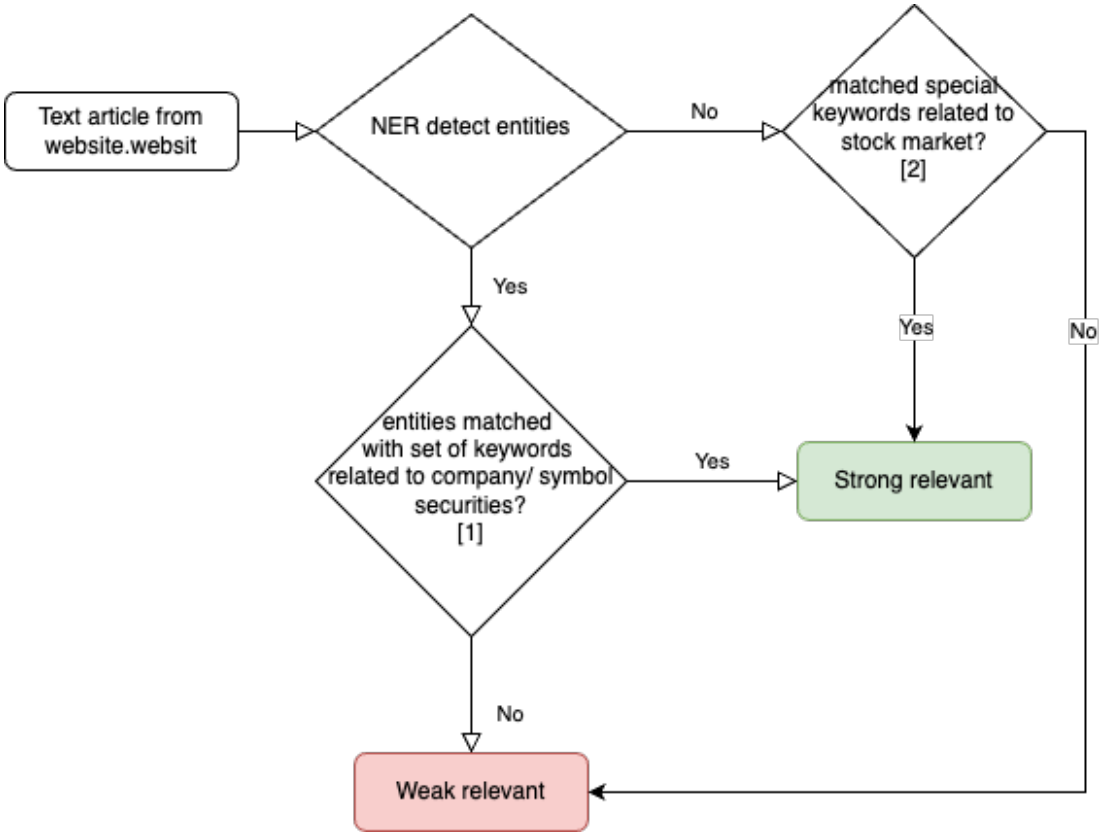

Figure 1. Flow chart for the Stock & Symbol Regconition via NER and special keywords

For example: if an entity passed to module [1] we will have a defined list of stock symbol, company name, people and places (ex: Vinamilk(VNM), Thegioididong(MWG),...) for fuzzy matching with the extracted entity. On the orther side, if it does not contained entities, we also do fuzzy matching on special keywords such as "thị trường chứng khoán",

"ngành thép", "cổ phiếu", "nhóm mã tăng trần",... Otherwise, if there is no matched item, we consider it is not relevant or irrelevant news.

Classifier Modeling: Models based on deep learning have repeatedly surpassed the performance of conventional machine learning techniques in various text categorization activities. Such tasks encompass analyzing sentiments, classifying news content, responding to questions, and inferring meanings in natural language, demonstrating deep learning's superior ability to interpret intricate patterns in extensive datasets (Minaee, 2021).

Train a machine learning model to classify news articles as relevant or irrelevant to stocks. You can use supervised learning with labeled data to build this classification model (Tun, 2021).

After the full text is extracted from the HTML by the preprocessing module, the content is then processed by the Stock Company and Symbol Recognition module. This module categorizes articles into two classes: strongly or weakly related to trading activities. We employ the BERT architecture to train this labeled dataset, an approach that has been validated for its efficacy in enhancing sequence tagging tasks in the Vietnamese language (Tran, 2020). Furthermore, we utilize the pre-trained viBERT models from prior work to bolster our experimental modeling efforts.

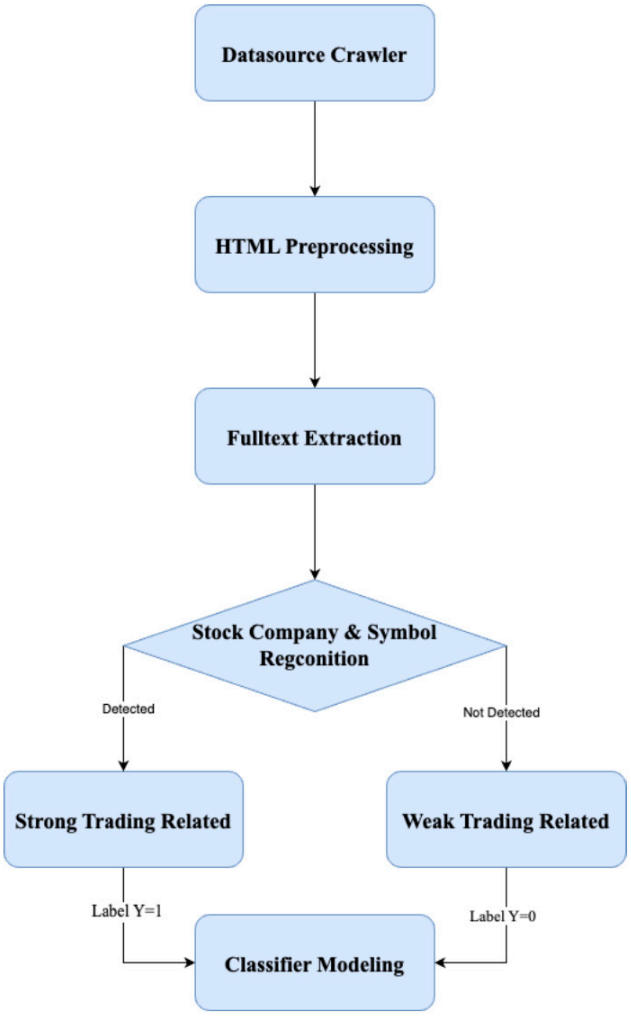

Figure 2. Full flow for data automation labeling and classification modeling

4. Results and discussion

4.1. Data Explorer

We amassed a comprehensive dataset comprising over 40,000 news articles throughout 2023. These articles were selectively sourced from the stock and economy sections of prominent financial websites, including vnEconomy, Cafef, Vietstock, and vietnambiz, among others. The distribution of articles across these sources is detailed in Figure 2.

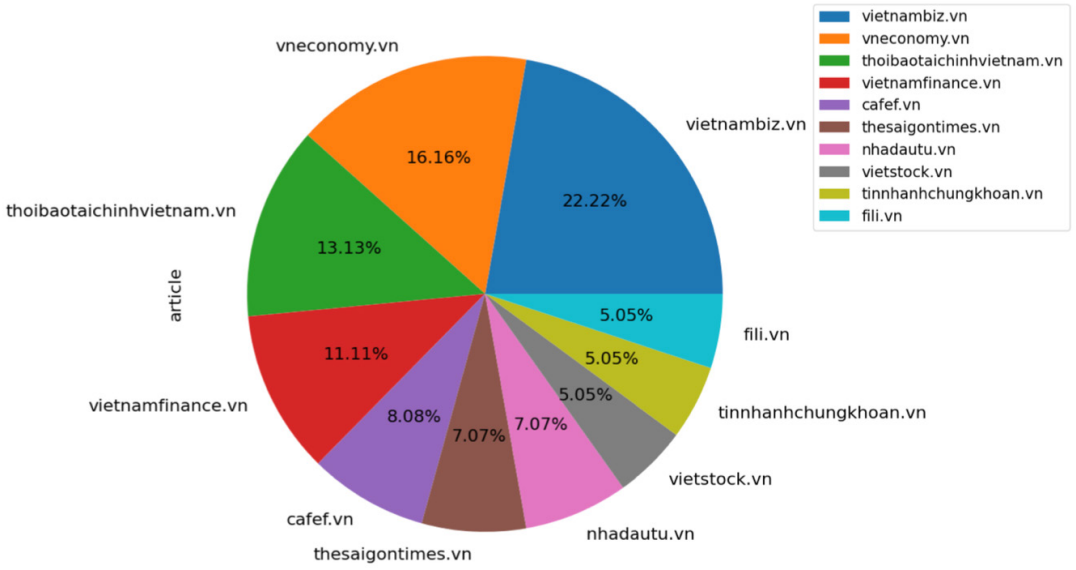

Figure 3. Data source description for the news classification model

We have included a comprehensive list of stock symbols from the three main exchanges in the Vietnamese market: HOSE, HNX, and UpCom. For example, from the HOSE exchange, we have HPG (Hoa Phat Group) and VNM (Vietnam Dairy Products Joint Stock Company). From HNX, one of the symbols is PLX (Vietnam National Petroleum Group). Additionally, from UpCom, we have ABI (Agriculture Bank Insurance Joint-Stock Corporation), among others. This dataset will serve as the input for training models in stock company and symbol recognition. The objective is to identify the names and symbols of companies listed on the stock market within a given text. Such recognition aids in filtering financial texts, studying market movements, and monitoring investor sentiment. To enhance the dataset's utility, we employed a named entity model, matched listed companies, and filtered stock domain keywords. This approach enabled the automatic tagging of news articles as related or unrelated to the stock market. Consequently, we gathered pertinent information about the companies traded on these exchanges. This intensive data collection process lays the groundwork for our future analysis and research into stock market dynamics and financial news sentiment analysis.

4.2. Classification with BERT

BERT, an acronym for Bidirectional Encoder Representations from Transformers, marks a notable progress in natural language processing (NLP) technology. This model, unveiled by Devlin in 2018, is engineered to develop deep bidirectional representations by simultaneously considering the context to the left and right of a word. Such a method provides the model with a deeper comprehension of language context

compared to the earlier models that only analyzed unidirectional context. Built upon the principles of a bidirectional Transformer encoder, which itself is an extension of the Transformer framework introduced by Vaswani in 2017, BERT's uniqueness stems from its foundational pre-training. This involves the model learning from a vast amount of untagged textual data through activities like masked language processing and predicting the following sentence. The benefit of this pre-training phase is that BERT acquires a comprehensive language understanding, which can then be refined with a single additional layer for various specific applications, including but not limited to, analyzing sentiments or answering questions. BERT's superiority has been validated across numerous standard datasets, underscoring its transformative impact on the development of NLP techniques

***Catastrophic Forgetting*

The phenomenon of "catastrophic forgetting" occurs in machine learning when a model loses previously acquired knowledge upon training with new data. To counteract this issue, particularly when training BERT models, a reduced learning rate is advisable. A learning rate of 2e-5 (0.00002) has been identified as effective in mitigating catastrophic forgetting, as meticulously determined in a study focused on fine-tuning BERT (Sun, 2019).

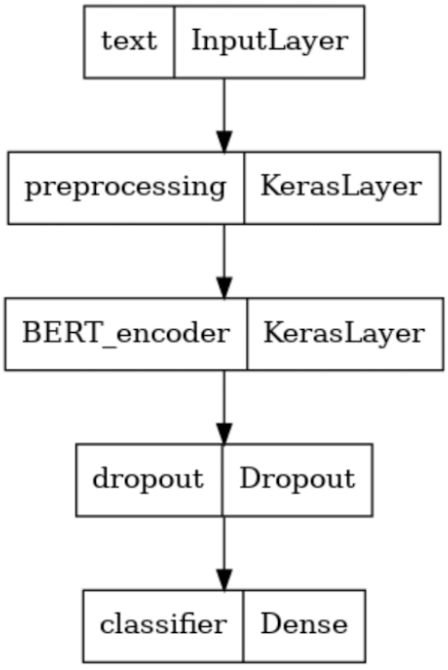

Figure 4.Introduce an example model structure of BERT classifier

Source: Complied from the https://www.tensorflow.org/

Figure 4 likely illustrates the typical workflow in a BERT-based text classification model. Here's a simplified description of the process:

1. **Raw Text Input:** The model starts with raw text data, which could be anything from sentences to full documents.
2. **Pre-processing:** This step involves cleaning and preparing the text for the model. It may include tokenization, where the text is split into tokens (words or sub-words), and encoding, where tokens are converted into numerical representations.

3. **BERT Encoder:** Following text pre-processing, it is submitted to the BERT encoder. BERT, an abbreviation for Bidirectional Encoder Representations from Transformers, analyzes each token in conjunction with every other token, generating a detailed, context-aware embedding for each.
4. **Dropout:** Subsequent to the encoding process, a dropout layer is incorporated. As a regularization method, dropout randomly zeroes out a portion of the input units during each training update, aiding in the reduction of overfitting.

Classifier (Dense Layer): Finally, the output from the dropout layer is fed into a dense layer, which is a fully connected neural network layer that outputs the probabilities of the different classes. For text classification, this would be the probabilities of each category the text might belong to.

We applied the architecture in this experiment with the standard hyperparameters of 12 layers, 768 hidden units, and 12 heads (Tran, 2020). Optimizer: Adam with betas=(0.9,0.999) and epsilon=1e-08.

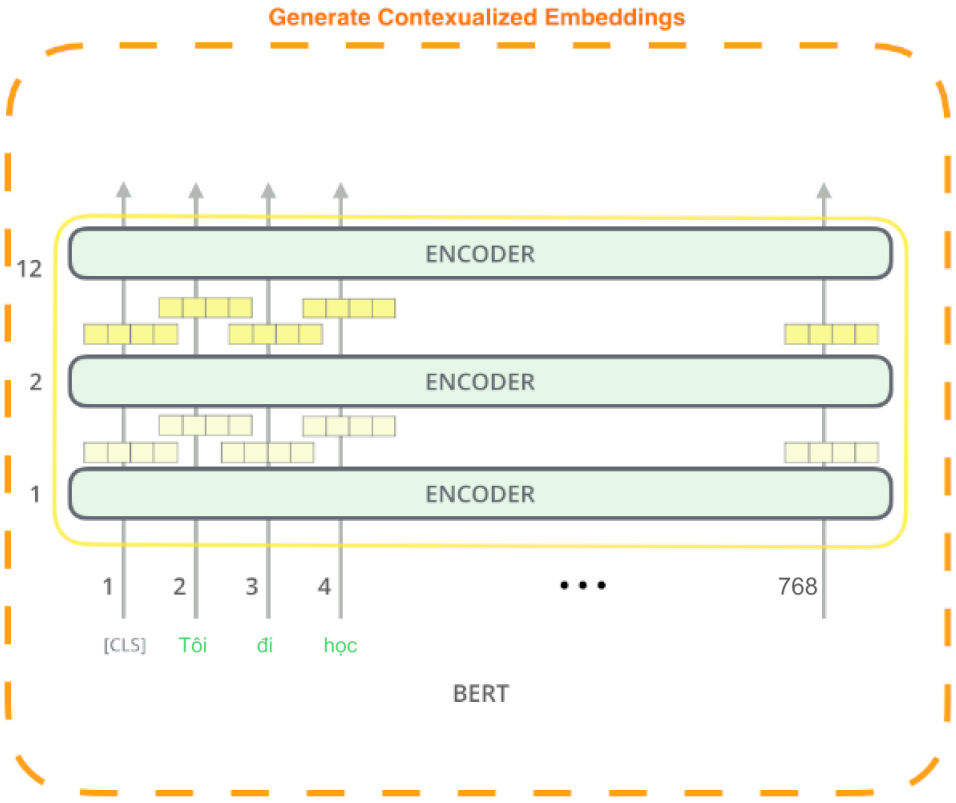

Figure 5. Our embedding BERT architecture

Source: Complied from the authors and the paper (Tran, 2020)

4.3. Experiment Evaluation

From a collection of 40,000 articles sourced from the Internet, we employed a news filter to identify content that mentioned stock codes, listed companies on the exchanges, or specialized stock market terminology. Subsequently, the data was bifurcated into two categories: articles with content matching

the filter criteria were deemed strongly related to stock trading, while the rest were classified as less relevant. Following this segmentation, our training dataset was composed of 40% of the articles labeled as '1 (indicating a strong relation to stocks), and the remaining 60% were labeled as '0' (indicating a weaker relation to stock trading)."

Our research endeavors to streamline the creation of high-quality datasets for machine learning applications. To this end, we amassed evaluation data meticulously labeled by trading experts. We applied our proposed method to this data without resorting to any tuning techniques such as parameter optimization, regularization, or dropout. The results were noteworthy: our method attained an overall accuracy of over (96%). The test dataset, rigorously evaluated by domain expert investors, formed the foundation of our analysis. This dataset was furnished by (STAG Vietnam, 2024).

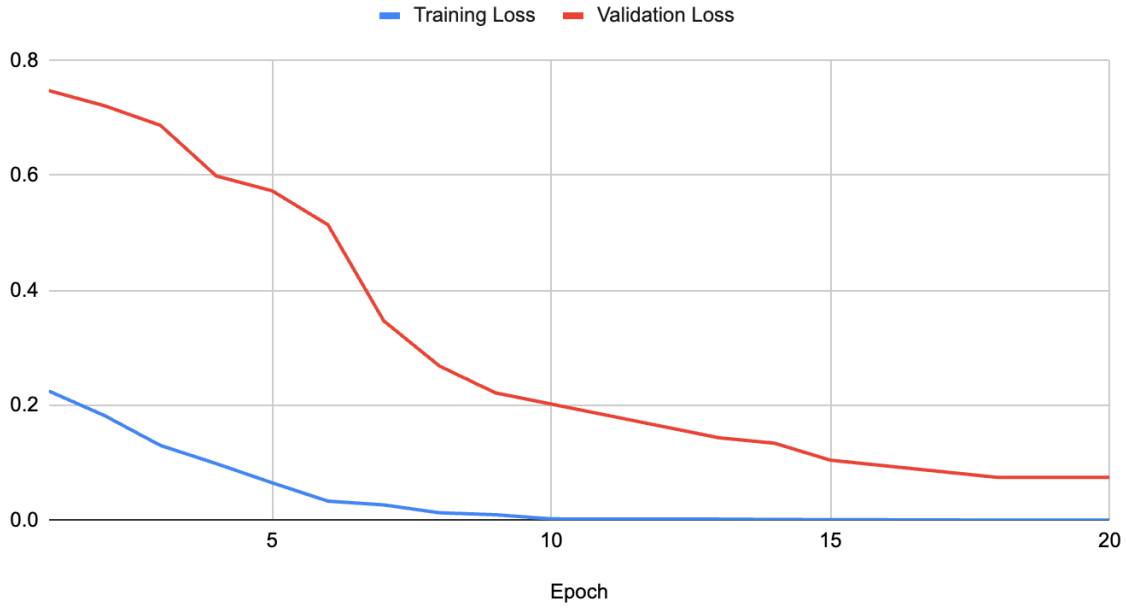

Figure 6. The training loss and validation loss of the training process

Upon arriving at the 16th epoch, both the training and validation losses demonstrate a plateau, indicating that the model may have achieved a state of optimal learning.

In this study, the computational resources were limited to a single CPU, specifically an Intel i5 10th generation, complemented by 16 GB of RAM and a 512 GB SSD. Additionally, a single GPU, the NVIDIA GeForce RTX 3060 with 12 GB of VRAM, was utilized.

5. Conclusions

This research addresses the challenge of finding relevant stock-related information among the massive of news and articles from Vietnamese economic websites and blogs, which also contain entertainment news, miscellaneous topics, and advertisements. These unrelated contents hinder the analysis of stock movements by genuine investors and analysts (Boudoukh, 2013). To solve this

problem, we propose an efficient method for automatic stock news categorization, based on a deep learning classifier model with BERT architecture in Vietnamese language (viBERT) (Tran, 2020). Our method aims to facilitate the access and reading of intensive stock related/ trading news for interested users.

The extension of this research will integrate sentiment analysis with stock return predictions, utilizing the viBERT model among other features. Sentiment analysis employs natural language processing to discern the opinions, emotions, and attitudes conveyed by authors or speakers within textual or spoken data. This technique is instrumental for predicting stock returns as it encapsulates market sentiment and public perception regarding companies, sectors, or events that influence stock prices (Makrehchi, 2013).

Acknowledgement

We would like to express our sincere appreciation to Ho Chi Minh University of Banking and the State Bank of Vietnam for their valuable guidance and support throughout this research project. We are also grateful to STAG Vietnam for their generous provision of the market data that enabled us to conduct our analysis. This research would not have been possible without their assistance and cooperation.

Reference

Allen, D. E. (2019). Daily market news sentiment and stock prices. *Applied Economics, 51*(30), 3212-3235. https://doi.org/10.1080/00036846.2018.1564115

Blum, A. A. (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the eleventh annual conference on Computational learning theory*, (pp. 92-100).

Boudoukh, J. A. (2013). *Which news moves stock prices? A textual analysis.* National Bureau of Economic Research.

Devlin, J. A.-W. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duong, D., Nguyen, T., & Dang, M. (2016, January). Stock market prediction using financial news articles on Ho Chi Minh Stock Exchange. In *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication* (pp. 1-6).

Francisco Viveros -Jiménez, M. A.-P.-A.-D. (2018). Improving the boilerpipe algorithm for boilerplate removal in news articles using html tree structure. *Computacion y Sistemas, 22*, 483-489.

Gidofalvi, G. A. (2001). Using news articles to predict stock price movements. *Department of computer science and engineering, university of california, san diego.* https://people.kth.se/~gyozo/docs/financial-prediction.pdf

Khan, W., Ghazanfar, M., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. (2020). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, *13*, 3433-3456. https://doi.org/10.1007/s12652-020-01839-w

Kohlschütter, C. (2022). *Boilerpipe*. Retrieved 01 2024, from Boilerpipe: https://github.com/kohlschutter/boilerpipe

Lison, P. A. (2020). Named entity recognition without labelled data: A weak supervision approach. *arXiv preprint arXiv:2004.14723*.

Makrehchi, M. A. (2013). Stock prediction using event-based sentiment analysis. *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, *1*, 337-342.

Minaee, S. A. (2021). Deep learning--based text classification: a comprehensive review. *ACM computing surveys (CSUR), 54*, 1-40.

Nguyen Van, P. (2015). *A good news or bad news has greater impact on the Vietnamese stock market?* (No. 61194). University Library of Munich, Germany.

Office, G. S. (2023). *Socio-economic situation report in the first quarter of 2023*. https://www.gso.gov.vn/en/highlight/2023/07/socio-economic-situation-report-in-the-first-quarter-of-2023/

Qing Li, T. W. (2014). The effect of news and public mood on stock movements. *Information Sciences*, 826-840.

Sun, C. A. (2019). How to fine-tune bert for text classification? *Chinese Computational Linguistics: 18th China National Conference, CCL 2019* (pp. 194-206). Kunming, China: Springer.

Sun, Y. M. (2018). A novel stock recommendation system using Guba sentiment analysis. *Personal and Ubiquitous Computing, 22*, 575-587.

Tran Duc Anh, N. S. (2023). *Stock Market Outlook 2024.* KB Securities Vietnam, Macro & Strategy. KB Securities Vietnam.

Tran, T. O. (2020). Improving sequence tagging for Vietnamese text using transformer-based neural models. In *Proceedings of the 34th Pacific Asia conference on language, information and computation* (pp. 13-20).

Triguero, I. A. (2015). Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems, 42*, 245--284.

Tun, N. A. (2021). Stock article title sentiment-based classification using PhoBERT. In *CEUR Workshop Proceedings* (Vol. 3026, pp. 225-233).

Usmani, S. A. (2021). News sensitive stock market prediction: literature review and suggestions. *PeerJ Computer Science, 7*, e490.

Van de Kauter, M. a. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, 4999-5010.

Ashish, V. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*, 1706.03762.

Villamil, L. A. (2023). Improved stock price movement classification using news articles based on embeddings and label smoothing. *arXiv preprint arXiv:2301.10458*.

Wu, H., Liu, Y., & Wang, J. (2020). Review of text classification methods on deep learning. *Computers, Materials and Continua*, *63*(3), 1309-1321. https://doi.org/10.32604/cmc.2020.010172

Xiaodong Li, H. X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems, 69*, 14-23.

Zhu, X. (2005). Semi-Supervised Learning Literature Survey. *World*, *10*. http://www2.denizyuret.com/ref/zhu/ssl_survey.pdf