Towards machine learning based text categorization in the financial domain

Frederic Voigt
University of the West of Scotland / Hamburg University of Applied Sciences
Department of Computer Science
Hamburg, Germany
b01742821@studentmail.uws.ac.uk

Jose Alcarez Calero
University of the West of Scotland
School of Computing, Engineering & Physical Sciences
Paisley, Scotland
jose.alcaraz-calero@uws.ac.uk

Keshav Dahal
University of the West of Scotland
School of Computing, Engineering & Physical Sciences
Paisley, Scotland
keshav.dahal@uws.ac.uk

Qi Wang
University of the West of Scotland
School of Computing, Engineering & Physical Sciences
Paisley, Scotland
qi.wang@uws.ac.uk

Kai von Luck
Hamburg University of Applied Sciences
Department of Computer Science
Hamburg, Germany
kai.vonluck@haw-hamburg.de

Peer Stelldinger
Hamburg University of Applied Sciences
Department of Computer Science
Hamburg, Germany
peer.stelldinger@haw-hamburg.de

***Abstract*—Despite the widespread research on text categorization in various Natural Language Processing (NLP) domains, there exists a noticeable void concerning its application to financial data. This study addresses this gap by employing pre-trained Bidirectional Encoder Representations from Transformers (BERT) models, fine-tuned specifically for the financial domain, to categorize newspaper articles focusing on financial topics. This is the first time that the dataset presented in this paper has been used. Further we evaluate the efficacy of established models in sentiment prediction using these rather long texts. Finally, we delve into the intricacies of company-specific sentiment and relevance prediction within these articles, acknowledging the prevalence of multiple companies being mentioned in one article, thus contributing to a more nuanced understanding of text analysis in the financial sector.**

***Index Terms*—natural language processing, stock price prediction, text categorization, finance, fundamental analysis**

I. Introduction

The utilization of machine learning within the stock price prediction (SPP) domain can be delineated into two primary categories. The first category, quantitative analysis, as explored in works such as [1], employs algorithmic models to forecast future stock prices based on historical ones [2]. The second category, fundamental analysis, extends the scope of data analysis beyond historical stock prices to include diverse sources of information such as corporate annual reports, news articles, or social media content [3]. In this context, NLP techniques are frequently utilized to systematically analyze and interpret the vast amounts of textual data, facilitating the extraction of meaningful insights that can impact financial decision-making processes. The impact of news on stock prices has been stated [4], underscoring the relevance of external information in financial markets. Furthermore, Huan, Wang, and Yang [5] highlight a critical challenge in this domain: SPP and Stock Movement Prediction (SMP) tasks are inherently complex due to the influence of information external to the analyzed texts. This factor can severely limit the accuracy of NLP models in making precise predictions, as these models may not account for critical non-textual influences on stock prices. Therefore a significant number of models prioritize sentiment analysis over the direct forecasting of stock prices or their movements derived from textual data.

Transformer models have become a cornerstone in the field of NLP and have found extensive applications in the financial sector. In the NLP domain, generic models often exhibit diminished performance when applied to finance-specific tasks, necessitating tailored approaches for optimal effectiveness [6]. Among these models, the BERT [7] model is particularly renowned for its popularity in processing financial texts. However, the deployment of Transformer models is constrained by their computational demands, with time complexity of $\mathcal{O}(n^2 \cdot k)$ and space complexity of $\mathcal{O}(n^2 + n \cdot k)$ [8], limiting inputs to a relatively small number of tokens. Numerous models have been tailored to handle shorter texts, such as news headlines, microblogs like StockTwits, or social media posts on platforms like X (formerly known as Twitter), as exemplified by [9], [10], and [11]. However, fewer studies, like the one conducted by Wang et al. [12], delve into longer texts. Apart from technical constraints, performance considerations also come into play, as articulated by Radinsky, Davidovich, and Markovitch [13], who assert that utilizing only news titles (instead of the full texts) yields superior results. Popular datasets in this domain include the SEMEVAL 2017 TASK 5 dataset [11] and the Financial Phrase Bank dataset [14], which contains expert-labeled sentences. The absence of models designed to handle entire news texts suggests that state-of-

979-8-3503- 8788-9/24/$31.00 © 2024 IEEE

the-art models may not perform as effectively on these longer texts compared to the previously mentioned shorter ones.

In addition to the scarcity of models designed to analyze entire (news) texts for classification, text categorization remains a relatively underexplored domain within NLP applied to finance. This underdevelopment likely stems from the dearth of available datasets tailored to this specific area of study. To the best of our knowledge, the FinBERT model, as presented by Huan, Wang, and Yang [5], is the only financial domain focused work designed to classify texts directly into categories such as environmental, social, and governance (ESG) texts. In contrast, other methodologies, such as the ECHO-GL approach developed by Liu et al. [15], incorporate textual topic classification as an intermediate step within a broader SMP process. The categorization of texts holds significant importance for two primary reasons. Firstly, within numerous investment firms, the manual execution of fundamental analyses by domain experts remains a common practice. In this context, the automation of pre-categorization processes stands to significantly enhance workflow efficiency, thereby conserving financial resources. Secondly, within the realm of machine learning, categorization could earn a pivotal role in the model pipeline. This is particularly evident as numerous models and methodologies are crafted, trained, and optimized for specific markets, such as those delineated in [16] or [17], as well as specific industries such as [18], [19], [17]. Furthermore, individual assets like Tesla [20] and Apple [21] are subject to tailored models. Moreover, cryptocurrencies have emerged as a prominent focal point for asset-specific modeling efforts, evidenced by publications such as [22] or [23] (Bitcoin).

The dearth of publicly accessible categorization datasets can be attributed, arguably, to the considerable expense associated with labeling the data. Fortunately, through the generous support and access to resources provided by Alpha Vantage (AV), a stock data supplier, we have been granted the privilege of accessing a diverse array of labeled online newspaper articles. These articles are not only annotated with sentiment analysis labels but also categorized into one of fifteen distinct topics, as illustrated in Figure 1. Moreover, the dataset offers a categorization of both the relevance and sentiment respective to all companies mentioned in the article. The aspect of articles mentioning multiple companies has been underscored in previous works such as [9], [4], and [10]. Furthermore, [4] states that the majority of companies mentioned within a text may not directly contribute to the sentiment prediction task. From an intuitive standpoint, predicting an overall sentiment becomes meaningless when the text discusses various companies, each potentially expressing distinct sentiments.

II. Related Work

As previously elucidated, a plethora of models have been developed to undertake sentiment prediction within financially pertinent texts, exemplified by [11]. However, a significant gap persists in the realm of text categorization, particularly within the financial domain, where models are conspicuously sparse. Notwithstanding this gap, within the broader domain of NLP, the subject remains a focal point of extensive research interest, as exemplified, for instance, by the work of Sun et al. [24]. In the domain of finance, the FinBERT model developed by Huan, Wang, and Yang stands out as one of the few models known to engage in text categorization. It merits acknowledgment that an extension has been published, broadening its classification scope to encompass nine categories, thereby introducing additional dimensions such as “Human Capital” or “Product Liability”. The FinBERT Model is distinguished for its pre-training from scratch, employing a diverse array of financial texts, totaling 4.9 billion tokens. Supplementing the model’s training, a specialized vocabulary, FinVocab, has been curated. To validate its efficacy, FinBERT undergoes fine-tuning and validation on three distinct sentiment classification datasets sourced from the financial domain, namely, the Financial Phrase Bank dataset, AnalystTone dataset, and FiQA dataset [25].

Another model, also called FinBERT, has been proposed by Araci [26]. To avoid confusions, we refer to it as “Araci-FinBERT”. However, in [25] it has been stated, that Araci-FinBERT shows inferior performance in comparison to FinBERT. The Araci-FinBERT model adopts a similar approach, utilizing the BERT architecture, further pre-training it on financial corpora and downstream tasks such as sentiment prediction within the financial domain. The Araci-FinBERT model diverges from the FinBERT model in that it has not undergone pre-training from scratch. The model introduced by Liu et al. [27], also referred to as “FinBERT”, lacks publicly available code, to the best of our knowledge, thus rendering it untestable. Additionally, it is remarkable that this model is also newly pretrained. In contrast, the FinBERT models accessible to the public have been scrutinized alongside the conventional BERT model, serving as the foundational comparison framework for the model employed in this study. BELT [10] from Dong et al., ECHO-GL [15] from Liu et al. and the work of Duan et al. [28] are the, to the best of our knowledge, only approaches to include a company-specific sentiment and/or relevance prediction mechanism in the pipeline.

III. Alpha Ventage Financial Data Corpus

Our dataset originates from Alpha Vantage[1], which generously granted us access to their API for our research endeavors. Within this platform, the Market News and Sentiment API offers a collection of pertinent articles covering financial news across $\psi = 15$ distinct topics. The Alpha Vantage API facilitates text search functionality for any given stock ticker. Initially, our data extraction focuses on companies listed in the S&P 500 index, which collectively represent approximately 40% of the total stock market capitalization [2]. Additionally, we extract data from the thirty companies listed on the DAX, the primary German stock index. The resulting dataset provides a

[1]https://www.alphavantage.co/

[2]Global Market https://www.sifma.org/resources/research/research-quarterly-equities/ valued at 109 trillion, with the S&P 500 constituting roughly 42 trillion Dollar https://www.slickcharts.com/sp500/marketcap

compilation of URLs to online news articles pertinent to each of the 530 companies under consideration. To download this data, we utilize the newspaper3k [3] library.

The dataset comprises URLs sourced from a variety of financially oriented websites, with specific distribution details outlined in Figure 1. The API establishes a maximum of 1000 articles per company within the period spanning from March 1, 2022, to February 10, 2024. The commencement of this period was determined empirically, marked by the earliest instance of article availability. The specified limit of 1000 articles per request was never met. Of the $265K$ URLs, $195,299$ were successfully downloaded; however, a substantial portion of these texts proved unsuitable due to impediments such as anti-web scraping measures or excessive access restrictions imposed by websites. An inherent constraint of BERT lies in its restricted capacity to process a limited number of tokens. Consequently, we retained only those data points that remained viable after tokenization, adhering to the token limit l. Following the exclusion of non-compliant websites, contentless pages, and overly lengthy articles, a total of $N = 2,556$ data points persisted. This represents approximately 50% of the size of the Financial Phrase Bank dataset, exceeding the dataset size employed in ESG classification as reported in [5] by 10%, and approximately matching the scale of the SEMEVAL 2017 TASK 5 dataset. For our analysis, we specifically utilize the title information within the text body. The title is demarcated with an additional token ([TITLE]), and it is incorporated as the inaugural line of the text.

Each newspaper article is construed as a labeled example, denoted by i, defined within the feature space $x_i \in V^l$, where $l = 512$ represents the standard maximum token length in numerous BERT models. We posit that the topic dataset $T = \{(x_i, Y_i^T)\}_{i=1}^N$ encompasses N observed samples generated through the AV-API, denoted as $(x_i, Y_i^T) \sim P_T(X, Y)$. Each $Y_i^T \subset 2^L \setminus \{\emptyset, L\}$ delineates the relevance classes of each x_i with respect to each of the ψ topics, employing the label set $L = \{l_j\}_{j=1}^{\psi}$, where $\forall l_j \in \{0, 1\}$ holds true. We assign $l_j = 1$ to indicate the relevance of an article to a specific topic j, and $l_j = 0$ to denote missing relevance. Thus, topic prediction manifests as a binary multi-label classification problem. The distribution and nomenclature of the topics are visualized in Figure 1.

Moreover, we establish the sentiment set $S = \{(x_i, Y_i^S)\}_{i=1}^N$, where $Y_i^S \in \mathbb{S}$ with $\mathbb{S} \in \{0, 1\}^{\theta_S}$ and $\theta_S = 3$. Our sentiment prediction model utilizes three classes, mirroring the approach of FinBERT, which also operates with three sentiment labels. These classes correspond to values between -1 and 1 in the original AV distribution.

For the sentiments, we establish an additional training set. Currently the negative sentiments constitute 11.5% of the data, neutral sentiments 62.3%, and positive sentiments 26.2%. The balanced sentiment set, denoted as $\hat{S} = \bigcup_{i=1}^{\theta_S} A(i) \subsetneq S$ assures $\forall i, j \in \mathbb{N} < \theta_S, |A(i)| = |A(j)|$ holds true with $A(j) \subseteq \{(x_i, Y_i^S) | Y_i^S[j] = 1\}$. Finally, $|\hat{S}| = 3 \cdot 294$ holds true, approximately matching the size of the 10%-Tone dataset introduced in [5]. Subsequently, we employ this dataset for model training, as outlined in Section V, to assess whether the learning process benefits from an equally distributed training data.

[3]https://newspaper.readthedocs.io/en/latest/

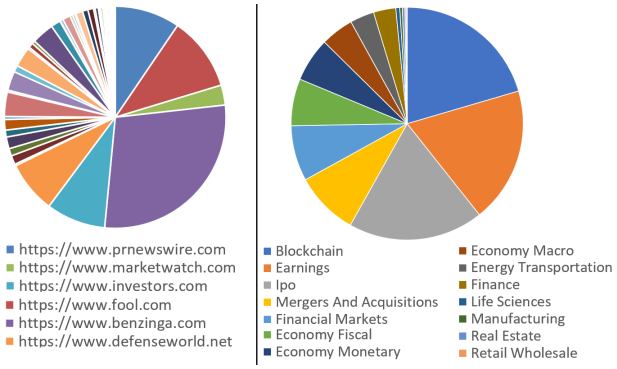

Fig. 1. (Left) Distribution of the 77 URLs. The six most frequent ones are listed below. (Right) Distribution of the ψ topic labels. The hardly recognizable proportions are "Life Science", "Real Estate" and "Manufacturing". It is important to acknowledge that multiple topics may be present within each article.

The dataset also provides sentiment and relevance scores for various publicly traded companies corresponding to each x_i. This enables the generation of information pertaining to over 8000 companies. While we provide some preliminary results in Section VIII, we defer a detailed investigation of the model's performance to future research endeavors.

IV. Model

We define the model $F_T : x_i \to 2^L \setminus \{\emptyset, L\}$ for the topic classification and $F_S : x_i \to \mathbb{S}$ for the sentiment classification. As the basis for both we choose pre-trained BERT models $F(x_i)[1] = h_{\text{CLS}}$ with $h_{\text{CLS}} \in \mathbb{R}^\eta$. The embedding size of the BERT model as well as the size of the attention-layers (inner layers) is set to $\eta = 768$ (having the default BERT model size as in [7]). In the majority of BERT models, this classification token ([CLS]-token) h_{CLS} functions as a condensed representation of the entire input text, as learned by the BERT model. The token undergoes additional processing through the BERT pooling layer before being passed into one of two layers, contingent upon the specific task being addressed. The linear sentiment layer employs the Softmax activation function to assign a probability to each of the three sentiment classes for x_i. On the other hand, the linear topic layer W_T is responsible for predicting the relevance Y^T across the $\psi = 15$ topics. Here, a Sigmoid activation function is utilized to ensure that each of the ψ outputs falls within the range of zero to one.

a) Loss Calculation: We establish two distinct loss functions, denoted as $\mathcal{L}_T(\hat{Y}^T, Y^T)$ employed for topic prediction utilizing binary Cross Entropy, and

$$\mathcal{L}_S = -\sum_{c=1}^{\theta_S} \alpha[c] \cdot Y^S[c] \cdot \log(\hat{Y}^S[c]) \tag{1}$$

with $\alpha[c] \propto |\{(x_i, Y_i^S) \in S \mid Y_i^S[c] = 1\}|^{-1}$ for the sentiment prediction as proposed in [29]. This formulation is intended to mitigate the dataset's imbalance within the sentiment dataset.

V. Training and Experiments

Following the methodology outlined in [5], our training regimen involves utilizing a training set comprising 81% of the data to train the model parameters, while a validation set, constituting 9% of the data, is employed for hyper-parameter optimization. The remaining 10% of the dataset is reserved for assessing model performance during testing. To ensure robustness and reliability of our results, we adopt a 10-fold cross-validation strategy, repeating each experiment 10 times. This approach guarantees that every data point has been included in the test set at least once. Hyperparameter tuning was conducted through an exhaustive grid search (untabulated). Ultimately, the in [7] recommended learning rate of 2×10^{-5} proved optimal. For the batch size, we found that a value of 32 yielded the best results. We conduct fine-tuning over 15 epochs, halting training upon observing a degradation in performance on the validation set or an increase in loss value (early stopping). As demonstrated in Section VI, the fifteen-epoch limit is never reached, as performance deteriorates well before reaching these thresholds. Initialization of the BERT model involves utilizing weights from various sources: the FinBERT model [5] (both pre-trained and fine-tuned on sentiment prediction, FLS (Forward-looking statements), or ESG/ESG-9 classification tasks), the Araci-FinBERT model [26], and the default pre-trained BERT model from [7].

VI. Results

The test set accuracy results pertaining to topic prediction are detailed in Table I. Notably, the Araci-FinBERT model exhibits the highest performance among the compared models. Comparatively, FinBERT models by Huang, Wang, and Yang demonstrate slightly inferior performance, trailing both the Araci-FinBERT model and the default BERT model, which lacks additional fine-tuning. The discrepancies in accuracy among the models are marginal, with differences typically observed within a few misclassified samples. It is remarkable that the FinBERT models developed by Huang, Wang, and Yang cease training after the initial epoch, due to deteriorating performance thereafter.

TABLE I
MEAN 10 RUN PERFORMANCE OF THE DIFFERENT BERT MODEL VARIANTS ON THE TOPIC CLASSIFICATION TASK. THE COLUMN "EPOCH" SHOWS IN WHICH EPOCH THE TRAINING HAS BEEN STOPPED (EACH OF THE 10 RUNS STOPPED AT THE SAME EPOCH ACROSS ALL MODELS).

Experiment	Accuracy	F1-Score	Epoch
FinBERT-tone	0.790	0.764	1
FinBERT-ESG	0.791	0.765	1
FinBERT-ESG-9	0.788	0.761	1
FinBERT-FLS	0.791	0.766	1
FinBERT-pretrained	0.784	0.756	1
Araci-FinBERT	**0.826**	0.820	3
BERT-pretrain	0.821	0.815	3

TABLE II
PERFORMANCE OF THE DIFFERENT BERT MODEL VARIANTS ON THE SENTIMENT CLASSIFICATION TASK. FOR THE "EPOCH" COLUMN THE MEDIAN VALUE HAS BEEN TAKEN.

Experiment	Accuracy	F1-Score	Epoch
FinBERT-tone	**0.902**	0.859	1
FinBERT-ESG	0.882	0.834	7
FinBERT-ESG-9	0.790	0.795	6
FinBERT-FLS	0.850	0.831	3
FinBERT-pretrained	0.737	0.704	8
Araci-FinBERT	0.815	0.792	4
BERT-pretrain	0.737	0.704	7

TABLE III
PERFORMANCE OF THE DIFFERENT BERT MODEL VARIANTS ON THE SENTIMENT CLASSIFICATION TASK USING $\hat{S}$ AS THE DATASET. FOR THE "EPOCH" COLUMN THE MEDIAN VALUE HAS BEEN TAKEN.

Experiment	Accuracy	F1-Score	Epoch
FinBERT-tone	**0.867**	0.851	1
FinBERT-ESG	0.777	0.750	5
FinBERT-ESG-9	0.759	0.729	6
FinBERT-FLS	0.775	0.747	9
FinBERT-pretrained	0.686	0.647	9
Araci-FinBERT	0.764	0.735	5
BERT-pretrain	0.720	0.685	5

The results pertaining to sentiment prediction are delineated in Table II. The FinBERT-tone model, already fine-tuned for sentiment prediction, exhibits superior performance. However, akin to the findings in topic prediction, the performance tends to decline beyond the first epoch of training. Remarkably, the performance of the FinBERT-tone model surpasses the benchmark accuracy of 88.2% reported in [5]. This phenomenon may be attributed in part to the dataset's magnitude, as emphasized by Huan, Wang, and Yang. Nonetheless, it is significant that the default BERT model's performance lags behind its counterpart reported in [5] by approximately 3%, a number similarly observed in the pre-trained FinBERT model.

Lastly, as depicted in Table III, it is evident that the performance metrics on a balanced dataset exhibit a decline of approximately 6%. Despite this reduction, the performance remains reasonably robust and can therefore not attributed to a dataset bias.

To delve into the model's internal mechanisms, we have depicted in Figure 2 the classification of a sample text into various topics using an adapted Grad-CAM [30] relevance visualization algorithm, elucidated further in Section VIII. The text discusses "Beyond Meat", a company grappling with significant challenges as it undergoes a transition from the period of IPO [4] excitement to a phase focused on ensuring its survival. Its stock price has plummeted due to financial concerns and a decrease in demand for plant-based foods, despite forging partnerships with major retailers and emphasizing cost-cutting measures and expansion into the European market. The text was disseminated by the AV-API for Yum! Brands,

[4]IPO = Initial Public Offering, is the process by which a private company offers stocks to the public for the first time.

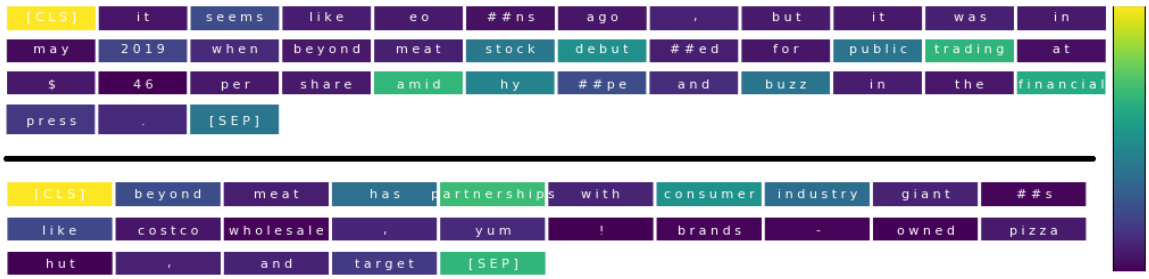

Fig. 2. Relevance visualization of an example of a sentence. The legend on the right-hand side denotes color significance, with high relevance represented at the top and low relevance at the bottom. The two sentences could not be entered together in the model, which is why they both contain a [CLS]-token and a [SEP]-token. Consequently, color scaling presented, is serving as a relative measure applicable within each sentence but not across the two sentences.

Inc., the world's largest foodservice group. In principle, this visualization illustrates the relevance of each token for the final classification. The article corresponds to the categories "IPO", "Retail & Wholesale", "Financial Markets", "Manufacturing", and "Earnings". As the output position of the [CLS]-token is passed into W_T, making it the single basis for the model's classification, the [CLS]-token exhibits the highest relevance in the visualization. Additionally, the [SEP]-token position likely holds significant relevance, as the model must attend to this token to effectively disregard outputs occurring after its position. Of particular importance is the elevated relevance attributed to phrases such as "stock debut" and "public trading", which encapsulate the essence of an "IPO" without explicit mention of the term itself. Remarkably, this relevance persists even amidst interruptions across other tokens. Additionally, the term "Consumer Industry" holds significant prominence, potentially aiding categorization within "Manufacturing" or even "Retail & Wholesale". Furthermore, mentions of "Financial Press" and "partnership" have also been underscored, indicating potential relevance for categorization within "Financial Markets". The term "Wholesale" is hardly considered relevant by the model even thought x_i is in the category "Retail & Wholesale".

VII. Conclusion

In summary, our experimentation involved evaluating various BERT models for sentiment prediction on financial news articles. The majority of these models demonstrated consistent performance, thereby validating their efficacy across diverse sentiment classifications within the financial domain. Moreover, commendable performance was attained for topic classification, a relatively novel task within the financial sector. Importantly, our findings highlight that the FinBERT model, when trained from scratch on financial data, exhibited the lowest performance across all variants. This observation suggests that retraining on financial data may not significantly enhance performance for both sentiment analysis and topic classification tasks on news article texts. Furthermore, it appears crucial that models exhibit continued performance improvement beyond the first training epoch, a characteristic lacking in the FinBERT model. This phenomenon may indicate that models, having already undergone training on financial data, reach a plateau of overfitting. Lastly, leveraging Grad-CAM inspired visualization techniques, we elucidated the robust internal mechanisms of the models for topic classification, revealing their adeptness at assigning phrases to respective topics within the financial domain.

VIII. Future Work

Future endeavors should prioritize the expansion and diversification of datasets tailored for topic classification within the realm of finance. This involves augmenting datasets to encompass a broader spectrum of topics, encompassing both additional and varied categories. Furthermore, efforts should extend towards diversifying more detailed sub-categories, such as various manufacturing sectors (e.g., automotive, airlines, ships, and consumer goods). Additionally, there exists ample opportunity to explore and include other market domains, including those within developing countries, national markets, niche markets, and beyond. Such diversification efforts will not only enhance the comprehensiveness of the datasets but also facilitate more robust training and testing of models within finance-related classification tasks. In the future, refining the model could involve additional enhancements, such as computing relevance and sentiment scores for specific companies, as outlined in Section I and Section III. Preliminary tests conducted in this regard have revealed particular challenges, especially concerning the weighting of the dataset. For instance, achieving an Accuracy of approximately 10% necessitates companies to appear at least 100 times, a performance that improves to around 38% when companies appear over 1000 times in the dataset. For future investigations, alternative techniques such as stop-word removal, TF-IDF methodologies, or summary generation could be explored to augment the dataset size. Additionally, broadening the scope to encompass additional companies could substantially enrich the corpus, facilitating more comprehensive analyses and yielding further insights.

Acknowledgment

The textual data as well as the labels presented in this paper were collected courtesy of research access kindly provided by Alpha Vantage [5]. We used the chatGPT AI [6] to improve the

[5]https://www.alphavantage.co
[6]https://chat.openai.com/

text in all sections of this work.

APPENDIX

We adapt the Grad-CAM visualization technique from Chefer et al. [31]. For implementation details, the interested reader is referred to the original literature. We adapt (and modify) the algorithm (following the notation in [31]). Each of the i input tokens is assigned a relevance $\hat{R}[i]$ with $\hat{R} \in \mathbb{R}^l$. This is calculated as

$$\hat{R}[i] = \frac{1}{2}\sum_{j=1}^{l} R[i,j] + \frac{1}{2}\sum_{v=1}^{l} R[v,i] \tag{2}$$

with

$$R = \bar{A} \text{ and } \triangledown A := \frac{\partial \sum_{c \in C} Y^T[c]}{\partial A} \tag{3}$$

and $C \subset \mathbb{N}$ containing all the indices of the topic classes to predict the relevance for.

REFERENCES

[1] F. Voigt, K. Von Luck, and P. Stelldinger, "Assessment of the Applicability of Large Language Models for Quantitative Stock Price Prediction," in *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '24, (New York, NY, USA), p. 293–302, Association for Computing Machinery, 2024.

[2] R. DeFusco, D. McLeavey, J. Pinto, and D. Runkle, *Quantitative Investment Analysis*. 2015. John Wiley Sons. (Cited on pages 1 and 3).

[3] A. S. Wafi, H. Hassan, and A. Mabrouk, "Fundamental Analysis Models in Financial Markets – Review Study," *Procedia Economics and Finance*, vol. 30, pp. 939–947, 2015. IISES 3rd and 4th Economics and Finance Conference.

[4] L. Yang, Z. Zhang, S. Xiong, L. Wei, J. Ng, L. Xu, and R. Dong, "Explainable Text-Driven Neural Network for Stock Prediction," in *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 441–445, 2018.

[5] A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A Large Language Model for Extracting Information from Financial Text*," *Contemporary Accounting Research*, vol. 40, no. 2, pp. 806–841, 2023.

[6] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. Chitkushev, and D. Trajanov, "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers," *IEEE Access*, vol. PP, pp. 1–1, 07 2020.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), pp. 4171–4186, Association for Computational Linguistics, 2019.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[9] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep Learning for Event-Driven Stock Prediction," in *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, p. 2327–2333, AAAI Press, 2015.

[10] Y. Dong, D. Yan, A. I. Almudaifer, S. Yan, Z. Jiang, and Y. Zhou, "BELT: A Pipeline for Stock Price Prediction Using News," in *2020 IEEE International Conference on Big Data (Big Data)*, pp. 1137–1146, 2020.

[11] K. Cortis, A. Freitas, T. Daudert, M. Huerlimann, M. Zarrouk, S. Handschuh, and B. Davis, "SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, and D. Jurgens, eds.), (Vancouver, Canada), pp. 519–535, Association for Computational Linguistics, Aug. 2017.

[12] G. Wang, T. Wang, B. Wang, D. Sambasivan, Z. Zhang, H. Zheng, B. Zhao, and S. Barbara, "Crowds on Wall Street: Extracting Value from Collaborative Investing Platforms," 03 2015.

[13] K. Radinsky, S. Davidovich, and S. Markovitch, "Learning causality for news events prediction," in *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, (New York, NY, USA), p. 909–918, Association for Computing Machinery, 2012.

[14] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," *Journal of the Association for Information Science and Technology*, vol. 65, 2014.

[15] M. Liu, M. Zhu, X. Wang, G. Ma, J. Yin, and X. Zheng, "ECHO-GL: Earnings Calls-Driven Heterogeneous Graph Learning for Stock Movement Prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 13972–13980, Mar. 2024.

[16] M. Leippold, Q. Wang, and W. Zhou, "Machine learning in the Chinese stock market," *Journal of Financial Economics*, vol. 145, no. 2, Part A, pp. 64–82, 2022.

[17] F. Onwuegbuche, J. Wafula, and J. Mung'atu, "Support Vector Machine for Sentiment Analysis of Nigerian Banks Financial Tweets," *Journal of Data Analysis and Information Processing*, vol. 07, pp. 153–173, 01 2019.

[18] P. Jariyapan, J. Singvejsakul, and C. Chaiboonsri, "A machine learning model for healthcare stocks forecasting in the us stock market during covid-19 period," *Journal of Physics: Conference Series*, vol. 2287, p. 012018, jun 2022.

[19] X. Xu, Y. Zhang, C. A. McGrory, J. Wu, and Y.-G. Wang, "Forecasting stock closing prices with an application to airline company data," *Data Science and Management*, vol. 6, no. 4, pp. 239–246, 2023.

[20] R. Fan, "Predictions of Tesla Stock Price based on Linear Regression, SVM, Random Forest, LSTM and ARIMA," *BCP Business Management*, vol. 44, pp. 422–431, 04 2023.

[21] P. Sonkiya, V. Bajpai, and A. Bansal, "Stock price prediction using BERT and GAN," 2021.

[22] V. Gurgul, S. Lessmann, and W. K. Härdle, "Forecasting Cryptocurrency Prices Using Deep Learning: Integrating Financial, Blockchain, and Text Data," 2023.

[23] J. Chen, "Analysis of Bitcoin Price Prediction Using Machine Learning," *Journal of Risk and Financial Management*, vol. 16, no. 1, 2023.

[24] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," in *Chinese Computational Linguistics* (M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, eds.), (Cham), pp. 194–206, Springer International Publishing, 2019.

[25] Y. Yang, M. C. S. Uy, and A. Huang, "FinBERT: A Pretrained Language Model for Financial Communications," *CoRR*, vol. abs/2006.08097, 2020.

[26] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," *CoRR*, vol. abs/1908.10063, 2019.

[27] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, "FinBERT: A pre-trained financial language representation model for financial text mining," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* (C. Bessiere, ed.), pp. 4513–4519, International Joint Conferences on Artificial Intelligence Organization, 7 2020. Special Track on AI in FinTech.

[28] J. Duan, Y. Zhang, X. Ding, C.-Y. Chang, and T. Liu, "Learning Target-Specific Representations of Financial News Documents For Cumulative Abnormal Return Prediction," in *Proceedings of the 27th International Conference on Computational Linguistics* (E. M. Bender, L. Derczynski, and P. Isabelle, eds.), (Santa Fe, New Mexico, USA), pp. 2823–2833, Association for Computational Linguistics, Aug. 2018.

[29] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," pp. 9260–9269, 06 2019.

[30] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization," *CoRR*, vol. abs/1610.02391, 2016.

[31] H. Chefer, S. Gur, and L. Wolf, "Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 387–396, 2021.