World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:17, No:2, 2023

A BERT-Based Model for Financial Social Media Sentiment Analysis

Josiel Delgadillo, Johnson Kinyua, Charles Mutigwe

***Abstract*—**The purpose of sentiment analysis is to determine the sentiment strength (e.g., positive, negative, neutral) from a textual source for good decision-making. Natural Language Processing (NLP) in domains such as financial markets requires knowledge of domain ontology, and pre-trained language models, such as BERT, have made significant breakthroughs in various NLP tasks by training on large-scale un-labeled generic corpora such as Wikipedia. However, sentiment analysis is a strong domain-dependent task. The rapid growth of social media has given users a platform to share their experiences and views about products, services, and processes, including financial markets. StockTwits and Twitter are social networks that allow the public to express their sentiments in real time. Hence, leveraging the success of unsupervised pre-training and a large amount of financial text available on social media platforms could potentially benefit a wide range of financial applications. This work is focused on sentiment analysis using social media text on platforms such as StockTwits and Twitter. To meet this need, SkyBERT, a domain-specific language model pre-trained and fine-tuned on financial corpora, has been developed. The results show that SkyBERT outperforms current state-of-the-art models in financial sentiment analysis. Extensive experimental results demonstrate the effectiveness and robustness of SkyBERT.

***Keywords*—**BERT, financial markets, Twitter, sentiment analysis.

I. Introduction

THE prediction of stock price movement is intrinsically a very challenging task because it depends on a multitude of factors such as fundamental indicators, technical indicators, political events, exchange rates, economic factors, etc. Fundamental indicators are data pertaining to a specified company including information about the intrinsic value based on economic variables such as the business model, company management, the balance sheet, cash flow, income statements, price to book value ratio, return on investment, etc. [23]. Technical indicators or data instead make use of price history to develop statistical indicators with the intent of analyzing trend and patterns in such history [23]. Technical analysis assumes that past values of the stock influence the future evolution of the market. However, the Efficient Market Hypothesis (EMH) states that stock price movements are largely driven by new information and follow a random walk pattern, and therefore do not follow any patterns or trends, and it is practically impossible to predict the future price movements based on the historical data [13], [14]. Recently, many researchers have proposed [41], [48], [49] various strategies that apply machine and deep learning algorithms in attempts to extract patterns in the way stock markets behave and respond to external factors. With the ubiquitous availability of social media services today, users post vast amounts of data expressing some judgment or sentiment about a topic or issue.

Stock trading can be a very dynamic and highly competitive activity in most financial markets. Traders use a combination of external information and internal company information to make trading decisions, and gaining an accurate vision of traders' opinions at scale can give a trader an advantage in making trading decisions. One such external source of information are microblogs posted on social media such as Twitter or StockTwits [43], [37], where users express their opinions on a variety of topics including stocks. NLP and machine learning techniques can be used to monitor market sentiment expressed in online news articles and/or social media posts in real-time, and use those sentiments as trading signals in buy or sell decisions. For example, if there is positive information about a particular company, it is expected that the company's stock price will increase, and vice versa. Bloomberg, the financial media company, has reported that trading sentiment portfolios outperform the benchmark index significantly [8]. Prior financial economics research also reports that news articles and social media sentiments could be used to predict market return and firm performance [40], [41].

In this research, the focus is on using financial social media posts and microblogs to develop models for the financial domain. This is done by leveraging the power of the BERT language model, pre-training a model, and then fine-tuning a model using various datasets as explained in section 3 using the approaches similar to BioBERT and FinBERT discussed in [11], [24], [4], and [25]. The model developed in this research, SkyBERT, is a domain-specific language representation model pre-trained on financial corpora.

The rest of this paper is organized as follows. Section II contains a literature review of sentiment analysis using NLP and deep neural networks. Section III explains the datasets and the methodology. Section IV contains the results, and analysis, and Section V provides a discussion. Section VI concludes the paper.

II. Related Work

Advances in machine learning have made financial text mining models in FinTech possible. NLP techniques can be used to better understand the large body of published financial text data. In particular, deep learning models are efficient and

Josiel Delgadillo and Johnson Kinyua are with The Pennsylvania State University, College of Information Sciences and Technology, University Park, PA, USA 16802 (e-mail: jkd5377@psu.edu, jdk450@psu.edu).

Charles Mutigwe is with Western New England University, 1215 Wilbraham Rd., Springfield, MA, USA 01119 (e-mail: charles.mutigwe@wne.edu).

effective for NLP tasks because they require relatively little feature engineering although they require a large amount of training data [38]. The focus of this research is the automated sentiment analysis of financial social media posts using NLP methods to classify such text as positive, negative, or neutral, in the financial domain. Recent unsupervised pre-training of language models on large corpora, such as BERT (Bidirectional Encoder Representations from Transformers) [12], ELMo [31], ULM Fit [17], XLNet, [49] and GPT [33] have significantly improved performance on many NLP tasks such as question answering, sentiment analysis, and language inference. These language models are trained on general domain corpora such as news articles and Wikipedia and are not suited to financial sentiment analysis tasks, because financial texts have a specialized language with a unique vocabulary. It is not difficult to fine-tune the language model using downstream tasks, but it has been shown [4] that pre-training a language model using a domain-specific corpus can further improve the task performance than fine-tuning the generic language model such as BERT. A few researchers have used this approach to create domain-specific BERT models as explained briefly in the following examples. BioBERT pre-trains a biomedical domain-specific language representation model using large-scale biomedical corpora [24], and ClinicalBERT applies the BERT model to clinical notes for hospital readmission prediction tasks [19]. To improve the of performance of downstream scientific NLP tasks, [5] developed SciBERT as a scientific domain-specific BERT model using a large multi-domain corpus of scientific publications. Zimbra et al. [47] have undertaken a thorough study and performance benchmark evaluation of Twitter sentiment analysis (TSA) systems across five domains (pharmaceuticals, retail, security, technology, and telecommunications) using 28 state-of-the-art systems. They report that the performance of these systems remains poor with reported tweet sentiment classification accuracies below 70%. The main challenges impacting the accuracy of TSA systems identified include the brevity of tweets (140 characters), novel language with Twitter-specific communication elements, strong sentiment class imbalance, and stream-based tweet generation. They further state that casual communications with frequent use of slang and acronyms prevalent in social media due to length restriction intensifies this communication behavior and promotes the development of a short novel language among Twitter users that makes it particularly difficult for TSA analyzers. Overall, in their studies, the systems performed poorly with a wide range of average classification accuracies ranging from 40% to 71%; while domain-specific approaches to sentiment analysis widely outperformed the general-purpose approaches with an improvement of 11%.

In the financial domain, some researchers have used similar approaches to develop several FinBERT models. The FinBERT model developed by Araci [4], [3] is pre-trained with a financial corpus and fine-tuned using a smaller financial dataset for sentiment classification in the financial domain. Their model is pre-trained with TRC2-financial which is a subset of the Reuters dataset [26], and the Financial PhraseBank dataset created by [27], used for fine-tuning the model. They implemented other pre-trained language models, LSTM, ULMFit, and ELMo for financial sentiment analysis for comparison with FinBERT. They reported that FinBERT increased the classification accuracy by 15% which is remarkable compared with these other models. Liu et al. [25] have developed a FinBERT model by starting with BERT and then taking it through six self-supervised pre-training tasks and then fine-tuning it with task-specific labeled financial data. They have reported that their FinBERT outperforms all previous state-of-the-art models in financial sentence boundary detection, financial sentiment analysis, and financial question-answering applications. During pre-training, FinBERT is simultaneously trained on a general corpus and a financial domain corpus, and during the fine-tuning phase, FinBERT is first initialized with the pre-trained parameters and is later fine-tuned on task-specific supervised data. The pre-training datasets used are: English Wikipedia and Books Corpus, Financial Web [7], [15], Yahoo! Finance [44], and Reddit Finance QA [34]. For fine-tuning, they use task-specific datasets depending on the intended use (financial sentence boundary detection, financial sentiment analysis, and financial question answering): FinSBD Shared Task dataset [30], Financial PhraseBank [27], FiQA Task 1, and FiQA Task2 [28]. Yang et al. have developed a FinBERT using a similar approach to the others [45]. They first compile a large financial domain corpus using corporate 10-K and 10-Q reports, earnings conference call transcripts, and analyst reports. They then used that to construct a financial vocabulary (FinVocab) for pre-training BERT. For fine-tuning, they used the Financial PhraseBank, FiQA Task 1 [28], and AnalystTone datasets [18] to develop different versions of FinBERT. They have reported that results show substantial improvement of FinBERT models over the generic BERT models. Desola et al. [10] have also developed three versions of FinBERT by pre-training BERT with 10-K SEC filings, and then tested using 10-Q SEC filings, and earnings call transcripts. They report that their FinBERT models outperform BERT in masked language model and next sentence prediction tasks. Through this research, a model called SkyBERT for sentiment analysis in the financial domain was developed starting with BERT, and then pre-training, and fine-tuning BERT using six financial domain-specific datasets as discussed in Section III.

III. Datasets & Methodology

A. Datasets

To develop the SkyBERT model as a domain-specific sentiment analyzer, six financial domain-specific datasets to pre-train, fine-tune, and test the model were used. The focus was to use social media posts and microblogs in Twitter posts that were related to the financial markets. The datasets used, and where in the model-building process they were used are discussed below.

i. Pre-training Dataset

Using a custom dictionary of financial terms and a custom

dictionary-based classifier, 150,000 financial news articles were extracted from the Thomson Reuters Text Research Collection (TRC2) corpus. The TRC2 corpus comprises 1,800,370 news stories covering the period from January 1, 2008, until February 28, 2009 [26]. Using the collection of financial news articles, a domain-specific dataset was created, referred to as FinTRC2. The FinTRC2 dataset is not labeled and was used to pre-train the model.

ii. Fine-tuning Dataset

The SkyBERT model was fine-tuned on the SSIX dataset. This dataset consists of 2,886 financial messages from StockTwits and Twitter with opinion targets [16]. The period of collection for this dataset was between October 2011 and June 2015. This dataset was annotated by financial experts using a scale of 1 to 7 for negative to positive sentiment at an entity level. This integer scale was eventually consolidated into a real number sentiment score in the [-1,1] range for each message. In the class of labeled datasets, the SSIX dataset provided a good sentiment class distribution at 23% for negative, 34% for neutral, and 44% for positive as shown in Fig. 1, and it was larger than the Taborda dataset, and was therefore selected for use in fine-tuning the model.

iii. Testing Datasets

For testing the SkyBERT models, as well as the other five comparative sentiment analyzers, the following datasets were used:

Fin-SoMe (FSM): The FSM dataset consists of 10,000 messages from StockTwits [6]. Chen et al. [6] did not provide a period for when these StockTwits messages were collected. Stocktwits is a microblogging platform that is popular with investors and traders, where they post stocktwits which consist of references to company stock symbols (so-called cashtags - a stock symbol preceded by "$", e.g., "$MSFT" for the company Microsof.), a short supporting text or references to a link or pictures. The FSM dataset is a gold standard that was annotated by experts [6] working in a bank's treasury marketing and risk management units. The market sentiment of each message in the FSM dataset was labeled as either bullish, bearish, or none.

Fin-Lin: The parent dataset (Fin-Lin) consists of 3,811 documents comprised of microblogs from StockTwits, news articles from Yahoo! News, financial reports for publicly traded companies, and analyst reports from July 1, 2018, to September 30, 2018 [9]. Only the StockTwits data were extracted from Fin-Lin to create the modified FinLin dataset, which consisted of 3,204 stock-twits/messages. Each message in the FinLin dataset was labeled with a numeric sentiment score in the [-1,1] range.

Sanders: The Sanders dataset consists of 5,512 tweets on four different topics (Apple, Google, Microsoft, and Twitter). This dataset is a gold standard with each tweet manually labeled by one annotator either positive, negative, neutral, or irrelevant with respect to the topic [35]. These tweets were collected between 2007 and 2011.

Taborda: The labelled stock market tweets dataset from Taborda et al. [39] consists of 1,300 tweets which were collected between April 9, 2020 and July 16, 2020, using the following Twitter tags as search parameter: #SPX500, #SP500, SPX500, SP500, $SPX, #stocks, $MSFT,$AAPL, $AMZN, $FB, $BBRK.B, $GOOG, $JNJ, $JPM, $V, $PG, $MA,$INTC $UNH, $BAC, $T, $HD, $XOM, $DIS, $VZ, $KO, $MRK, $CMCSA, $CVX, $PEP, $PFE. The tweets were manually annotated with positive, neutral, or negative sentiment classes [39]. The sentiment class distribution of the testing datasets is shown in Fig. 1. Given the imbalanced class distributions, the F1 scores rather than the accuracy as the primary metric to evaluate the models was used [32].

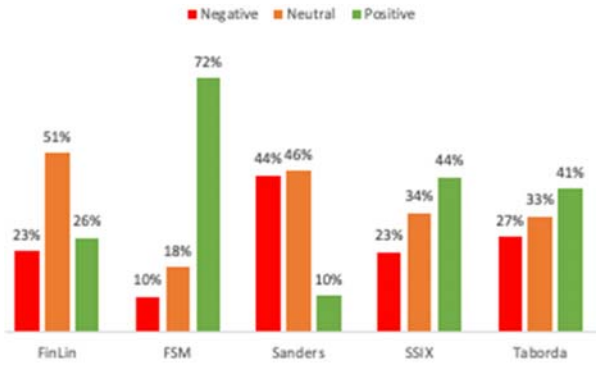

Fig. 1 Sentiment distribution of the fine-tuning and testing datasets

To test the performance of the model in a realistic scenario, it was crucial to ensure that the fine-tuning and testing datasets were independent. This also made sure that the SkyBERT model, like the five other external sentiment analyzers, was not trained or fine-tuned on the testing data. It was also important to ensure that the testing data were collected from different sources and at different times for the fine-tuning data sets, these constraints were designed to help remove any data source or temporal related biases.

B. Data Pre-processing

Each sentiment score in the SSIX and FL-ST datasets was converted into a categorical score. The messages that had a sentiment score in the [-1, -0.33] range was classified as negative, those with a score between (-0.333, 0.333) are classified as neutral, and finally, those scores between [0.333, 1] were classified as positive. For the FSM dataset, the sentiment labels were reclassified as positive for messages labeled as 'bullish', negative for 'bearish', and neutral for 'none'. There was no need to make any modifications to the class labels for the Sanders and Taborda datasets. The data in the SSIX dataset were shuffled before using it for fine-tuning. A custom shuffling approach was used similar to the one proposed by Nguyen et al. [29].

C. Methodology

The SkyBERT model was developed on the Amazon SageMaker platform using Python 3.6 and the following frameworks and libraries: TensorFlow, PyTorch, and Transformers. The model development process began with a $BERT_{BASE}$ model [12] and using a process similar to that discussed by Devlin et al. [12]. The Sky-BERT model was pre-trained using the FinTRC2 dataset. Next, the model was fine-

tuned using the SSIX dataset. The hyperparameters recommended by Devlin et al. [12] were used with a batch size of 32 and fine-tuning for 3 epochs with a learning rate of 2e-5. More experiments with other hyperparameters such as increasing batch size to maximize the number of training examples per epoch and increasing the epochs were conducted. To evaluate how the model performed on tweets and other microblogs in the financial domain, the model was tested using the FSM, Sanders, Taborda, and FinLin datasets. For external validation, performance evaluation was done with five other sentiment analyzers using the FSM, Sanders, Taborda, and FinLin datasets. For the five external sentiment analyzers, three academic sentiment analyzers were used: VADER [20], SentiStrength [42], and FinBERT [45], and two commercial ones: IBM Watson Natural Language Understanding (NLU) [22] and Amazon Comprehend [1]. VADER and SentiStrength are lexicon-based, while FinBERT, like SkyBERT, is a deep learning model that is based on transformers. The architectures of Watson NLU and Amazon Comprehend were not available. A local model of VADER was accessed using Python, and a local model of SentiStrength as a Java JAR file was used [36]. The Python APIs were used to remotely access the IBM Watson NLU [21], Amazon Comprehend [2], and FinBERT [46] sentiment analysis models.

IV. Results

The model development process required 19 epochs to reach an optimal result, as shown in Fig. 2. During fine-tuning, SkyBERT was able to produce healthy results when tracking training and validation accuracy and loss rather than demonstrating the model had underfit or overfit. SkyBERT required more epochs as compared to the 3 epochs recommended by Devlin et al. [12] and 3 by Araci for FinBERT [4]. Similarities were drawn with the learning rate of 2e-5 which was used by Devlin et al. [12] and Araci [4]; and experimentation with the other learning rates was used but did not prove to be a better model during testing. Dropout rate, batch size, scheduler, and weight decay were other aspects of the model that were experimented with during fine-tuning to ultimately find the best-performing model based on the weighted F1 score.

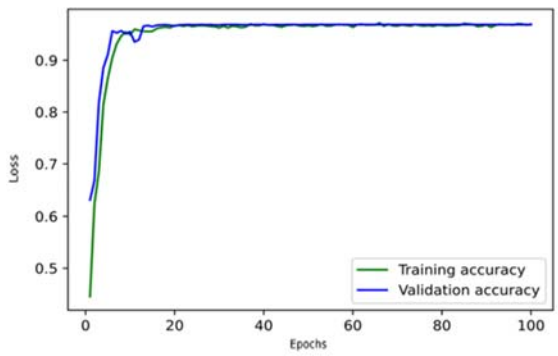

Fig. 2 SkyBERT training and validation accuracy

TABLE I
WEIGHTED F1 SCORES FOR THE SENTIMENT ANALYZERS

Model	FinLin	FSM	Sanders	Taborda	AVG
VADER	0.501	0.539	0.409	0.663	**0.525**
SkyBERT	0.473	0.527	0.417	0.485	**0.477**
Watson NLU	0.462	0.425	0.488	0.513	**0.472**
SentiStrength	0.476	0.459	0.385	0.475	**0.449**
FinBERT	0.545	0.402	0.341	0.438	**0.432**
Amazon Comprehend	0.481	0.312	0.491	0.385	**0.417**

TABLE II
MACRO F1 SCORES FOR THE SENTIMENT ANALYZERS

Model	FinLin	FSM	Sanders	Taborda	AVG
VADER	0.476	0.393	0.393	0.658	**0.465**
SkyBERT	0.515	0.406	0.357	0.485	**0.447**
Watson NLU	0.449	0.324	0.456	0.515	**0.436**
SentiStrength	0.426	0.364	0.369	0.479	**0.410**
FinBERT	0.487	0.350	0.331	0.4238	**0.398**
Amazon Comprehend	0.413	0.306	0.467	0.398	**0.396**

The training and validation loss characteristics are shown in Fig. 3. The weighted F1 scores shown Table I and the macro F1 scores shown in Table II were used to compare the performance of the six sentiment analyzers. The results shown in Table I indicate that the SkyBERT model outperformed all the other models, except VADER, based on the average (AVG) scores.

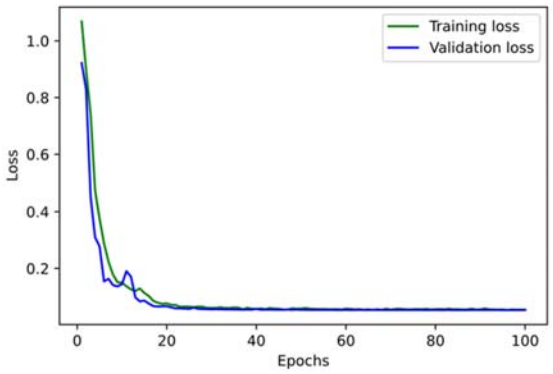

Fig. 3 SkyBERT training and validation loss

V. Discussion

Although the F1 score was used as the primary metric compared to Zimbra et al. [47] who used accuracy, it can be seen from Table III that all of the six sentiment analyzers that were studied would have accuracies in the 40% to 70% range. This shows that the performance of TSA in the financial domain (FTSA) is not significantly different from that of the other domains in Zimbra et al. [47]. This suggests that there is room for improvement and more research is needed to develop FTSA models with higher accuracy using the techniques discussed by Zimbra et al. [47].

The FTSA model closest to SkyBERT in terms of architecture is FinBERT [45]. The results show that SkyBERT outperformed FinBERT in 3 of the 4 tests. The conclusion is that this superior performance is due to the larger financial domain pre-training dataset that was used for SkyBERT and the fine-tuning that was done using only social media posts.

TABLE III
ACCURACY SCORES FOR THE SENTIMENT ANALYZERS

Model	FinLin	FSM	Sanders	Taborda	AVG
VADER	0.499	0.500	0.411	0.666	**0.513**
FinBERT	0.593	0.384	0.468	0.483	**0.482**
SkyBERT	0.510	0.434	0.417	0.480	**0.464**
Watson NLU	0.461	0.376	0.491	0.516	**0.461**
Amazon Comprehend	0.544	0.314	0.523	0.448	**0.457**
SentiStrength	0.497	0.418	0.412	0.496	**0.456**

SkyBERT lagged VADER in all but one test. According to Hutto and Gilbert, VADER uses a set of lexical features that are specifically attuned to sentiment in microblog-like contexts [20]. The experiments conducted here use a relatively small dataset (2,886 stocktwits and tweets) to fine-tune SkyBERT so that it becomes a sentiment-aware model. In the future, if these experiments are conducted using a larger and more diverse financial social media dataset to fine-tune SkyBERT, the expectation is that improved results will be obtained. The discussion above regarding SkyBERT and FinBERT seems to support this idea. While SentiStrength is also a lexical-based sentiment analyzer like VADER, it is not as optimized for sentiment detection in microblog-like contexts. As a result, SkyBERT outperformed SentiStrength in 3 out of 4 of the tests.

SkyBERT outperformed the two commercial sentiment analyzers, although the performance of Watson NLU was much closer to that of SkyBERT. The lack of transparency concerning the architecture and training of the commercial sentiment analyzers makes it difficult for us to posit reasons for their poor performance relative to SkyBERT and VADER.

VI. CONCLUSION

In this work, a BERT-based model called SkyBERT is proposed, which is a pre-trained language model for financial-task-oriented sentiment analysis. SkyBERT is pre-trained with FinTRC2, a financial domain corpus, which enabled the SkyBERT model to effectively capture language knowledge and semantic information in the financial domain; and then fine-tuned with the SSIX dataset enabling SkyBERT to understand financial domain knowledge embedded in social media posts in StockTwits and Twitter. Extensive experimental results demonstrated the effectiveness and robustness of SkyBERT compared with FinBERT [45] and other sentiment analyzers.

In future work, the plan is to fine-tune SkyBERT using a much larger financial social media dataset and to test SkyBERT sentiment analysis of stock tweets, and other sentiment analyzers used in this research on live trading decisions in the financial markets. With the release of SkyBERT later, it is hoped that practitioners and researchers can utilize SkyBERT for a wide range of applications where the prediction goes beyond sentiment analysis, such as financial-related outcomes such as stock returns, stock volatilities, and fraud.

ACKNOWLEDGMENTS

We would like to thank Amazon AWS for the summer 2021 grant that allowed us to port and complete this research on the SageMaker platform and for the support from the AWS US Higher Ed. team.

REFERENCES

[1] Amazon. (n.d.). Amazon Comprehend: Features. Retrieved June25, 2022 from https://aws.amazon.com/comprehend/features
[2] Amazon Web Services. (n.d.). Amazon Comprehend Developer Guide. Retrieved June 25, 2022 from https://docs.aws.amazon.com/comprehend/latest/dg/comprehend-dg.pdf.how-sentiment
[3] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. CoRR abs/1908.10063 (2019). arXiv: 1908.10063 http://arxiv.org/abs/1908.10063
[4] Doğu Tan Araci and Zulkuf Genc. July 31, 2020. FinBERT: Financial Sentiment Analysis with BERT. Prosus AI Tech Blog. Retrieved July 1, 2022 from https://medium.com/prosus-ai-tech-blog/finbert-financial-sentiment-analysis-with-bert-b277a3607101
[5] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pre-trained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371
[6] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Issues and Perspectives from 10,000 Annotated Financial Social Media Data. In Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, 6106–6110. https://aclanthology.org/2020.lrec-1.749
[7] Common Crawl. (n.d.). Retrieved November 30, 2021 from https://commoncrawl.org/
[8] X. Cui, D. Lam, and A. Verma.2016. Embedded Valuein Bloomberg News and Social Sentiment Data. Technical Report.
[9] Tobias Daudert. 2020. A Multi-Source Entity-Level Sentiment Corpus for the Financial Domain: The FinLin Corpus. CoRR abs/2003.04073 (2020). https://arxiv.org/abs/2003.04073
[10] Vinicio Desola, Kevin Hanna, and Pri Nonis. 2019. FinBERT: pre-trained model on SEC filings for financial natural language tasks. Technical Report. https://doi.org/10.13140/RG.2.2.19153.89442
[11] Jacob Devlin and Ming-Wei Chang. November 2, 2018. Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. Google AI Blog. Retrieved July 1, 2022 from https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html
[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423
[13] Eugene F. Fama. 1965. Random Walks in Stock Market Prices. Financial Analysts Journal 21, 5 (1965), 55–59. http://www.jstor.org/stable/4469865
[14] Eugene F. Fama. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance 25, 2 (1970), 383–417. http://www.jstor.org/stable/2325486
[15] FinancialWeb. (n.d.). Retrieved November 30,2021 from https://www.finweb.com/
[16] Thomas Gaillat, Manel Zarrouk, André Freitas, and Brian Davis. 2018. The SSIX Corpora: Three Gold Standard Corpora for Sentiment Analysis in English, Spanish and German Financial Microblogs. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1423
[17] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, 328–339. https://doi.org/10.18653/v1/P18-1031
[18] Allen H. Huang, Amy Y. Zang, and Rong Zheng. 2014. Evidence on the Information Content of Text in Analyst Reports. The Accounting Review 89, 6 (06 2014), 2151–2180. https://doi.org/10.2308/accr-50833
[19] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT:

Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 (2019).
[20] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media 8, 1 (May 2014), 216–225. https://ojs.aaai.org/index.php/ICWSM/article/view/14550
[21] IBM. (n. d.). IBM Cloud API Docs: Natural Language Understanding. Retrieved June 25, 2022 from https://cloud.ibm.com/apidocs/natural-language-understanding?code=python
[22] IBM. (n. d.). Watson Natural Language Understanding: Features. Retrieved June 25, 2022, from https://www.ibm.com/cloud/watson-natural-language-understanding/details
[23] Investopedia. (n.d.). Financial Terms Dictionary. Retrieved November30, 2021 from https://www.investopedia.com/financial-term-dictionary-4769738
[24] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (sep 2019). https://doi.org/10.1093/bioinformatics/btz682
[25] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. 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), Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4513–4519. https://doi.org/10.24963/ijcai.2020/622
[26] Reuters Ltd. 2004. Reuters Corpora (RCV1, RCV2, TRC2). National Institute of Standards and Technology. https://trec.nist.gov/data/reuters/reuters.html
[27] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius,and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology 65, 4 (2014), 782–796. https://doi.org/10.1002/asi.23062
[28] Financial Opinion Mining and Question Answering. 2017. https://sites.google.com/view/fiqa/
[29] Thao Truong Nguyen, François Trahay, Jens Domke, Aleksandr Drozd, Emil Vatai, Jianwei Liao, Mohamed Wahib, and Balazs Gerofi. 2022. Why globally reshuffle? Revisiting data shuffling in large scale deep learning. In IPDPS 2022: 36th International Parallel & Distributed Processing Symposium. IEEE, Lyon (virtual), France. https://hal.archives-ouvertes.fr/hal-03599740
[30] The First Workshop on Financial Technology and Natural Language Processing (FinNLP) with a Shared Task for Sentence Boundary Detection in PDF Noisy Text in the Financial Domain (FinSBD). (n. d.). https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp/
[31] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. https://doi.org/10.18653/v1/N18-1202
[32] Purva Huilgol. August 24, 2019. Accuracy vs. F1 score. Retrieved June 30, 2022, from https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2
[33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. OpenAI Blog 1, 8 (2019), 9.
[34] Reddit.(n.d.). Retrieved November 30, 2021 from https://www.reddit.com/
[35] Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. 2013. Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS Gold. In 1st International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013). http://oro.open.ac.uk/40660/
[36] SentiStrength. (n.d.). Retrieved June 25, 2022 from http://sentistrength.wlv.ac.uk/
[37] StockTwits, Inc. (n.d.). Retrieved November 30, 2021 from https://stocktwits.com/
[38] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In 2017 IEEE International Conference on Computer Vision (ICCV). 843–852. https://doi.org/10.1109/ICCV.2017.97
[39] Bruno Taborda, Anade Almeida, José Carlos Dias, Fernando Batista, and Ricardo Ribeiro. 2021. Stock Market Tweets Data. https://doi.org/10.21227/g8vy-5w61
[40] Paul C. Tetlock. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance 62, 3 (2007), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x
[41] Paul C. Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. More Than Words: Quantifying Language to Measure Firms' Fundamentals. The Journal of Finance 63, 3 (2008), 1437–1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x
[42] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, DiCai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61,12(2010),2544–2558. https://doi.org/10.1002/asi.21416
[43] Twitter, Inc. (n. d.). Retrieved November 30, 2021, from https://twitter.com/
[44] Yahoo! Finance. (n. d.). Retrieved November 30, 2021, from https://finance.yahoo.com/
[45] Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. FinBERT: A Pre-trained Language Model for Financial Communications. arXiv e-prints (June 2020). https://doi.org/10.48550/arXiv.1908.10063
[46] Yang, Y. (n. d.). FinBERT. Retrieved June 25, 2022 from https://github.com/yya518/FinBERT/
[47] David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2018. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. ACM Trans. Manage. Inf. Syst. 9, 2, Article 5 (Aug 2018), 29 pages. org/10.1145/3185045
[48] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Rsulan Salakhutdinov, and Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. 2019. arXiv:1906.08237. https://arxiv.org/abs/1906.08237
[49] Mojtaba nabipour, Pooyan Nayyeri, Hamed Jabani, Shahab S., and Amir Mosavi. Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis. (2020). DOI: 10.1109/ACCESS.2020.3015966. https://ieeexplore.ieee.org/abstract/document/9165760