Is Domain Adaptation Worth Your Investment? Comparing BERT and FinBERT on Financial Tasks

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, Chu-Ren Huang
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University,
Yuk Choi Road 11, Hung Hom, Kowloon, Hong Kong
{peng-bo.peng,emmanuele.chersoni,yu-yin.hsu,churen.huang}.polyu.edu.hk

Abstract

With the recent rise in popularity of Transformer models in Natural Language Processing, research efforts have been dedicated to the development of domain-adapted versions of BERT-like architectures.

In this study, we focus on FinBERT, a Transformer model trained on text from the financial domain. By comparing its performances with the original BERT on a wide variety of financial text processing tasks, we found continual pretraining from the original model to be the more beneficial option. Domain-specific pretraining from scratch, conversely, seems to be less effective.

1 Introduction

The Transformer architectures have taken the field of Natural Language Processing (NLP) by storm, leading to remarkable performance leaps in several tasks (Vaswani et al., 2017; Devlin et al., 2019).

The first-generation Transformers were mainly trained on general corpora, such as Wikipedia or Common Crawl. However, considering domain adaptations, many researchers have later injected domain-specific knowledge in such architectures, leading to the publication of Transformers trained on different types of in-domain text, e.g., scientific articles (Beltagy et al., 2019), biomedical text (Lee et al., 2020; Gu et al., 2020), clinical notes (Alsentzer et al., 2019), and patent corpora (Lee and Hsiang, 2020).

Since language technologies have seen increasingly frequent use in accounting and finance (Loughran and McDonald, 2016), it is not surprising that several attempts have been made to adapt Transformers to the financial domain (Araci, 2019; Yang et al., 2020; Liu et al., 2020).

In this study, we test the FinBERT model by Yang et al. (2020) on a variety of tasks in the field of financial NLP, including sentiment analysis, causality detection, numeral understanding, and numeral attachment, and we study the impact of different types of pretraining on the system performance. We obtained the best results with a FinBERT model with pretraining continuing from the original BERT and with the same general-domain vocabulary, while a model trained anew on financial corpora and with a domain-adapted vocabulary performed similarly to BERT Base.

2 Related Work

Although financial NLP is a relatively recent field, it already has an active research community, which has regularly introduced new shared tasks and benchmarks in recent years, e.g., sentence boundary detection in financial documents (Azzi et al., 2019; Wan et al., 2019; Au et al., 2021), hypernymy detection (El Maarouf et al., 2021; Mansar et al., 2021), document causality detection (Mariko et al., 2020), document structure extraction (Juge et al., 2019; Bentabet et al., 2020), and document summarization (Zheng et al., 2020). Given the success of Transformer models in general-domain NLP, it is not surprising that they are also a popular choice for many systems competing in financial tasks (Chen et al., 2020).

To adapt the original BERT to sentiment analysis in the financial domain, Araci (2019) was the first to propose a FinBERT model by further pretraining BERT Base on the financial subset of the Reuters TRC2 corpus. The evaluation, carried out on the Financial Phrase Bank (Malo et al., 2014) and the FiQA sentiment scoring dataset (Maia et al., 2018), demonstrated that FinBERT largely outperformed all the LSTM-based baselines and was slightly better than the original model.

The second FinBERT model, introduced by Yang et al. (2020), followed two different training strategies. The first version (FinBERT BaseVocab) was further pretrained from a BERT Base checkpoint on three financial corpora (i.e., the Corporate Reports 10-K & 10-Q from the Securities

Proceedings of the Third Workshop on Economics and Natural Language Processing, pages 37–44
Punta Cana and online, November 11.

Exchange Commission, [1] the Earnings Call Transcripts from the Seeking Alpha website, [2] and the Analyst Reports from the Investext database), and the second (FinBERT FinVocab) was trained afresh on the same three corpora but with new vocabulary specific to the financial domain, not inheriting it from the original BERT. They evaluated the models on the same sentiment analysis datasets, in conjunction with the opinion mining data from Huang et al. (2014), and reported improved performance over BERT Base, especially when using the FinBERT model with the domain-adapted vocabulary.

In this study, we chose to test the FinBERT system used in Yang et al. (2020), which has two publicly available versions, in order to directly compare the impact of the two different domain adaptation strategies and to evaluate them on more semantic tasks. The previous studies (Araci, 2019; Yang et al., 2020) focused their evaluation exclusively on sentiment analysis. However, sentiment analysis is a general task that is not necessarily ideal for observing the advantages of domain adaptation because the expressions of sentiment might not reflect the in-domain language. For example, in the biomedical domain, several tasks have recently been shown to benefit from training from scratch on an in-domain text and from a domain-specific vocabulary (Gu et al., 2020; Portelli et al., 2021). Therefore, besides sentiment analysis, we decided to evaluate our models on three semantic tasks that are more specific to the financial domain: document causality detection (Mariko et al., 2020), numeral understanding (Chen et al., 2019b), and numeral attachment (Chen et al., 2020).

3 Experimental Setting

3.1 Tasks and Datasets

The following section describes all the tasks related to the study, and the datasets to be evaluated. Descriptive statistics for the latter are provided in Table 1. More details about the class distributions are in Appendix A.

3.1.1 Sentiment Analysis

Sentiment Analysis stands out as one of the most popular tasks in NLP. To compare our models in the financial domain, we selected three different datasets. The Financial PhraseBank (Malo et al., 2014) is a standard dataset for sentiment classification composed of 4,840 sentences selected from financial news and annotated for Positive, Negative, and Neutral sentiment by 16 different annotators with experience in the financial domain. The dataset comes with the original annotations: for our study, we evaluated on a subset of 2,264 instances with at least 75% of annotator agreement.

We also used the FinTextSen dataset from SemEval 2017 Task 5 that dedicates itself to sentiment analysis on financial microblogs (Cortis et al., 2017). The dataset consists of 2,488 microblog messages retrieved from Twitter and StockTwits in March 2016. Each instance contains the following information: the message, a cashtag, and a sentiment score. The latter was originally a continuous score, but we used the dataset version by Daudert et al. (2018), who clustered the scores to obtain a 3-class annotation (Positive, Negative, and Neutral), to maintain consistency with the other sets.

Finally, the StockSen dataset (Xing et al., 2020) is composed of 20,675 financial tweets extracted from the StockTwits platform between June and August 2019, all of which were annotated with either Positive or Negative sentiments.

3.1.2 Financial Document Causality Detection

For Document Causality Detection, we used the dataset of the FinCausal shared task 2020 (Mariko et al., 2020). The dataset is made of texts extracted from a 2019 corpus of financial news provided by Qwan, with each instance annotated with binary labels to indicate whether it described a causal relation. For example, in (1), the italicized part was annotated as the cause for the fall of the GDP.

(1) *Things got worse when the Wall came down.* GDP fell 20% between 1988 and 1993.

We refer to the dataset for subtask 1, which is a simple binary classification task (class 1 if the text includes a causal relation and 0 otherwise).

3.1.3 Numeral Understanding

Understanding numerals is of key importance for the automatic processing of financial documents. In coincidence with the FinNum shared task, Chen et al. (2019b) released a microblogs dataset extracted from StockTwits, in which numerals are annotated with 7 high-level categories (i.e., **Monetary, Percentage, Option, Indicator, Temporal, Quantity, and Product/Version**) and 17 more

[1] https://www.sec.gov/edgar.shtml.
[2] https://seekingalpha.com/.

Datasets	Train	Dev	Test	Classes	Max Length
Financial Phrase Bank (Malo et al., 2014)	2,264	\	\	3	81
FinTextSen (Daudert et al., 2018)	2,488	\	\	3	476
StockSen (Xing et al., 2020)	14,457	6,218	\	2	370
Causality Detection (Mariko et al., 2020)	13,478	\	8,580	2	1,460
FinNum-1 subtask 1/2 (Chen et al., 2019b)	4,072	457	786	7/17	48
FinNum-2 (Chen et al., 2019a)	7,187	2,109	1,044	2	120

Table 1: Descriptive statistics for all the experimental datasets: train and test splits, classes, max text length.

fine-grained classes, which are sub-classes of the same categories. The labels have been identified based on the taxonomy by Chen et al. (2018), and the annotation was carried out by two domain experts. The dataset only includes examples on which the annotators reached an agreement. Examples (2a) and (2b) illustrate, respectively, the Monetary and the Product/Version category (the numeral expression to be classified is in bold).

(2) a. $FB (**110.20**) is starting to show some relative strength and signs of potential B/O on the daily.

b. iPhone **6** may not be as secure as Apple thought.. $AAPL

We address both the subtasks of FinNum (e.g., the 7-class and the 17-class classification tasks); that is, the tweets containing n financial numbers and the corresponding category labels will be copied n times. The details of the reconstructed data are also illustrated in Table 1.

3.1.4 Numeral Attachment

The numeral attachment task was introduced during the FinNum-2 competition (Chen et al., 2019a). The authors built a dataset of financial microblogs extracted from StockTwits, in which, given a target cashtag and a target numeral, a system predicts whether the numeral is attached to the cashtag. For example, in (3), the second numeral in the sentence is attached to the $NE cashtag, while the first one is not.

(3) **$NE**, last time oil was over *$65* you were close to *$8*.

Therefore, for each instance, the system must perform a binary classification task (i.e., 1 if the numeral is attached to the cashtag, and 0 otherwise).

3.2 Models

In this study, two baseline models were used. One is the **BERT Base** (Devlin et al., 2019), which consists of a series of stacked Transformer encoders. It was trained using both a masked language modeling objective and a next sentence prediction objective on a concatenation of the Books Corpus (Zhu et al., 2015) and the English version of Wikipedia. The other one is a traditional Support Vector Machine (SVM) baseline (Noble, 2006), where the input representation is the element-wise addition of the word vectors of each word in the sentence. We used the publicly available FastText vectors by Grave et al. (2018).

As for the FinBERT models, we used **FinBERT BaseVocab** (FV w/ BV) and **FinBERT FinVocab** (FB w/ FV) (Yang et al., 2020). The former was initialized from the original BERT Base (i.e., it also uses the same general-domain vocabulary) and then further pretrained on financial corpora, and the latter was trained afresh on financial corpora for 1M iterations and uses a domain-specific financial vocabulary.

Following the methodology by Devlin et al. (2019), all models used a linear layer with $softmax$ as a classification layer and the cross-entropy loss as a loss function. The texts were directly fed to the models after some simple preprocessing steps. For all models, we replaced the URLs with the special token [URL]. For the Numeral Understanding task, the texts and the target numbers were concatenated with the special token [SEP] after the tokenization. Finally, in the Numeral Attachment task, we followed Moreno et al. (2020) by adding the special tokens **£** and **§** to the beginning and the end of the *$cashtag*, and the target number, respectively.

3.3 Evaluation Metrics

All the models have been evaluated in terms of Macro F1-score and Micro F1-score. In this study,

Datasets	SVM		FB w/ BV		FB w/ FV		BERT Base	
	Micro-F1(%)	Macro-F1 (%)	Micro-F1(%)	Macro-F1 (%)	Micro-F1 (%)	Macro-F1 (%)	Micro-F1 (%)	Macro-F1 (%)
Financial Phrase Bank	61.62	33.55	**96.86±1.43**	**95.51±2.32**	96.69±1.1	95.39±1.54	96.60±1.06	95.15±1.52
FinTextSen	69.53	36.81	**84.48±2.34**	56.81±3.59	83.08±3.25	57.34±7.65	83.04±2.58	**60.83±10.94**
StockSen	73.21	42.90	**79.48±0.7**	**69.69±0.41**	76.37±0.57	69.38±0.75	78.72±0.92	68.78±0.48
Causality Detection	94.07	59.84	94.28±0.68	**79.79±0.81**	**94.51±0.31**	79.68±0.71	94.24±0.6	79.65±0.74
FinNum-1 subtask 1	63.27	34.69	**94.38±0.29**	**89.41±0.79**	93.51±0.72	87.11±1.07	94.07±0.48	88.04±1.39
FinNum-1 subtask 2	48.76	24.40	**88.84±0.51**	**80.71±0.82**	87.45±0.67	80.66±1.62	88.12±0.63	79.4±1.5
FinNum-2	82.69	51.64	85.67±0.55	67.56±2.12	**85.78±0.52**	**67.84±2.8**	85.07±0.44	66.51±1.91

Table 2: Comparative results in terms of Micro-F1 and Macro-F1 (top scores per dataset/metric are in **bold**), with standard deviations for the BERT models.

Datasets	FB w/ BV vs. BERT Base		FB w/ FV vs. BERT Base		FB w/ BV vs. FB w/ FV	
	Micro-F1(%)	Macro-F1(%)	Micro-F1(%)	Macro-F1(%)	Micro-F1(%)	Macro-F1(%)
Financial Phrase Bank	0.26	0.36	0.09	0.24	0.17	0.12
FinTextSen	1.44	-4.02	0.04	-3.49	1.4	-0.53
StockSen	0.76	0.91	-2.35	0.6	3.11	0.31
Causality Detection	0.04	0.14	0.27	0.03	-0.23	0.11
FinNum-1 subtask 1	0.31	1.37	-0.56	-0.93	0.87	2.3
FinNum-1 subtask 2	0.72	1.31	-0.67	1.26	1.39	0.05
FinNum-2	0.6	1.05	0.71	1.33	-0.11	-0.28
Sentiment Analysis	0.82	-0.92	-0.74	-0.88	1.56	-0.03
Numeral Understanding	0.52	1.34	-0.62	0.17	1.13	1.18

Table 3: Performance gaps for each dataset and metric. In the last two lines, we also report the aggregate performance for the group of sentiment analysis datasets (Financial Phrase Bank, FinTextSen and StockSen) and for the numeral understanding ones (FinNum-1 subtask 1 and 2).

the latter is equivalent to the traditional Accuracy metric, due to treating each task as a multi-class classification task. For the datasets without an official train-test split (e.g., FinTextSen and Financial Phrase Bank), we ran a 10-fold cross-validation and reported the average score. However, due to the instability of BERT fine-tuning on small datasets (Zhang et al., 2020), even the results of multiple runs on the same split may heavily fluctuate. Therefore, we reported the average scores after 10 runs, even for the datasets with an official train-test split.

4 Results and Discussion

The full results are shown in Table 2. Firstly, we observe that all the pretrained BERT models outperformed the SVM baseline in all the financial datasets. Secondly, many models reported large standard deviations on some of the datasets, especially the sentiment analysis ones. It can be observed that FinBERT BaseVocab reports the best performance in almost all the datasets, generally outperforming BERT Base. Excluding the FinTextSen dataset, in which BERT Base is the top-scoring model, FinBERT BaseVocab achieves an average increase of 0.85 of Macro F1-score on the other benchmarks. On the other hand, FinBERT FinVocab performed similarly to BERT Base, sometimes showing small improvements and sometimes lagging behind the original model. It achieved the top score only in the numeral attachment task and in causality detection, the latter only for the Micro-F1. [3] Moreover, the performance increase for FinBERT BaseVocab was more noticeable on the datasets on numerals, while the performances of FinBERT FinVocab were more irregular, performing slightly better than BERT Base and the BaseVocab model on FinNum2 (numeral attachment), but lagging behind both on FinNum subtask 1 (numeral understanding).

Table 3 summarizes the performance comparison between the Transformer models, where it can be seen that FinBert BaseVocab typically improves over the other models for both metrics (the FinTextSen dataset being the only exception). However, it should also be noticed that the differences between models are sometimes small compared to the standard deviations in Table 2, which invites to be cautious in drawing firm conclusions.

4.1 Error Analysis

We ran a qualitative error analysis of the instances that were misclassified by our models for the tasks of sentiment analysis, numeral attachment, and

[3] It should be pointed out that in the Causality data the class distribution is very unbalanced, with almost 93% of negative instances (see Appendix A), and thus Macro-F1 is a more reliable score.

Text instance	Task	Golden Label	Misclassified by
$AAPL Force in VWAP is strong with this one.....no break since it fell below.....awesome	StockSen	0	All
$GOOG $AMZN $FB Trump is not going to do anything to these companies. He wouldn t risk crashing the market before the election. That anti-trust talk is just smoke and mirrors.	StockSen	1	FB w/ BV
£$HMNY£ it's over. No one is going back. Once people get a deal. *§30§* years ago I sold Toyotas for full sticker only. The world changes!	FinNum-2	0	All
£$SPY£ Tax reform scam is code word for bailout. After *§8§* years, the CBs are still pumping. They want to transfer wealth. Don't let them.	FinNum-2	0	All
When they signed up in 2008, the government invested R52-million to fund the workers shares.	Causality Detection	0	All
The existing $500 - $600 billion of public support for agriculture must be redirected to more inclusive, resilient and low carbon production and innovative technologies and finance to enhance the resilience of small-scale producers.	Causality Detection	0	BertBase

Table 4: Error cases for different tasks, together with the right label and the models that misclassified the instance.

causality detection. Table 4 displays some of the examples that we extracted.

For Sentiment Analysis, we extracted some misclassified examples from StockSen and noticed that the polarity of some tweets is mistaken by the classifiers because of irony, such as the final exclamation *awesome* on the first row in Table 4. In some other cases, like the one on the second row, the words associated with a negative polarity (e.g., *risk, crashing*) might be misleading the systems, while the tweet is actually positive.

In the numeral attachment task, where the target cashtag is in bold, and the target numeral in italics, the models seem to experience problems in assigning the correct interpretations to numerals, especially when they appear in temporal adjuncts (e.g., the examples on the third and the fourth rows).

The error sources seem to be more varied and more difficult to identify in the causality detection task. However, we encountered a few cases like the examples on the fifth and the sixth rows, where a *to*-infinitive construction is used for expressing goals. Given the semantic similarity between cause and goal, it seems plausible that the construction has confused the classifiers, leading them to erroneously assign the instances to the positive class.

5 Conclusions

In this paper, we compared the original BERT model with the financially adapted models by Yang et al. (2020). Domain adaptation was generally confirmed to be beneficial and, unlike what has been recently observed in the biomedical domain (Gu et al., 2020; Portelli et al., 2021), the model benefiting from continuous pretraining from BERT Base showed more consistent improvements across tasks and datasets. This suggests that the models take advantage from exposure to financial text, but the tasks do not necessarily require a specialized vocabulary. On the negative side, fluctations in the results confirmed that there is some degree of instability in the fine-tuning of BERT-like models on relatively small datasets (Zhang et al., 2020).

In our future work, we plan to investigate also the contextualized embeddings produced by the domain-adapted Transformers. Word embeddings have been used in tasks with important applications in the financial domain, such as the identification of semantic relations (Chersoni et al., 2016; Xiang et al., 2020), which is useful for building domain ontologies (El Maarouf et al., 2021; Mansar et al., 2021; Chersoni and Huang, 2021), and the unsupervised detection of semantic changes in diachronic data, e.g., annual reports of traded companies (Giulianelli et al., 2020; Montariol et al., 2021; Masson and Montariol, 2021). In this perspective, a promising research direction would be to analyze how different domain adaptation strategies affect the quality of the embedding representations.

Acknowledgments

We would like to thank Tobias Daudert, Frank Xing, and Chung-Chi Chen for sharing their datasets with us, and the four anonymous reviewers for their insightful feedback. This research was made possible by the University Postdoc Matching Fund (W16H) and Project of Strategic Importance (ZE2J) at the Hong Kong Polytechnic University.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the NAACL Workshop on Clinical Natural Language Processing*.

Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063*.

Willy Au, Abderrahim Ait-Azzi, and Juyeon Kang. 2021. FinSBD-2021: The 3rd Shared Task on Structure Boundary Detection in Unstructured Text in the Financial Domain. In *Companion Proceedings of the Web Conference*.

Abderrahim Ait Azzi, Houda Bouamor, and Sira Ferradans. 2019. The FinSBD-2019 Shared Task: Sentence Boundary Detection in Pdf Noisy Text in the Financial Domain. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv preprint arXiv:1903.10676*.

Najah-Imane Bentabet, Rémi Juge, Ismail El Maarouf, Virginie Mouilleron, Dialekti Valsamou-Stanislawski, and Mahmoud El-Haj. 2020. The Financial Document Structure Extraction Shared Task (FinToc 2020). In *Proceedings of the Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019a. Numeral Attachment with Auxiliary Tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1161–1164.

Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. 2018. Numeral Understanding in Financial Tweets for Fine-grained Crowd-based Forecasting. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 136–143. IEEE.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019b. Overview of the NTCIR-14 FinNum Task: Fine-grained Numeral Understanding in Financial Social Media Data. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, pages 19–27.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2020. Overview of the NTCIR-15 FinNum-2 Task: Numeral Attachment in Financial Tweets. *Development*, 850(194):1–044.

Emmanuele Chersoni and Chu-Ren Huang. 2021. PolyU-CBS at the FinSim-2 Task: Combining Distributional, String-Based and Transformers-Based Features for Hypernymy Detection in the Financial Domain. In *Companion Proceedings of the Web Conference*, pages 316–319.

Emmanuele Chersoni, Giulia Rambelli, and Enrico Santus. 2016. CogALex-V Shared Task: ROOT18. In *Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon*.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 Task 5: Fine-grained Sentiment Analysis on Financial Microblogs and News. In *Proceedings of SemEval*.

Tobias Daudert, Paul Buitelaar, and Sapna Negi. 2018. Leveraging News Sentiment to Improve Microblog Sentiment Classification in the Financial Domain. In *Proceedings of the EMNLP Workshop on Economics and Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Ismail El Maarouf, Youness Mansar, Virginie Mouilleron, and Dialekti Valsamou-Stanislawski. 2021. The FinSim 2020 Shared Task: Learning Semantic Representations for the Financial Domain. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing*.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of ACL*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of LREC*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific Language Model Pretraining for Biomedical Natural Language Processing. *arXiv preprint arXiv:2007.15779*.

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):2151–2180.

Rémi Juge, Imane Bentabet, and Sira Ferradans. 2019. The FinToc-2019 Shared Task: Financial Document Structure Extraction. In *Proceedings of the Financial Narrative Processing Workshop*.

Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent Classification by Fine-tuning BERT Language Model. *World Patent Information*, 61:101965.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, 36(4):1234–1240.

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 IJCAI*.

Tim Loughran and Bill McDonald. 2016. Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, 54(4):1187–1230.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. In *Companion Proceedings of The Web Conference*, pages 1941–1942.

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):782–796.

Youness Mansar, Juyeon Kang, and Ismail El Maarouf. 2021. The FinSim-2 2021 Shared Task: Learning Semantic Similarities for the Financial Domain. In *Companion Proceedings of the Web Conference*, pages 288–292.

Dominique Mariko, Estelle Labidurie, Yagmur Ozturk, Hanna Abi Akl, and Hugues de Mazancourt. 2020. Data Processing and Annotation Schemes for FinCausal Shared Task. *arXiv preprint arXiv:2012.02498*.

Corentin Masson and Syrielle Montariol. 2021. Detecting Omissions of Risk Factors in Company Annual Reports. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing*.

Syrielle Montariol, Alexandre Allauzen, and Asanobu Kitamoto. 2021. Variations in Word Usage for the Financial Domain. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing*.

Jose G Moreno, Emanuela Boros, and Antoine Doucet. 2020. TLR at the NTCIR-15 FinNum-2 Task: Improving Text Classifiers for Numeral Attachment in Financial Social Data. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo*, pages 8–11.

William S Noble. 2006. What is a Support Vector Machine? *Nature Biotechnology*, 24(12):1565–1567.

Beatrice Portelli, Edoardo Lenzi, Emmanuele Chersoni, Giuseppe Serra, and Enrico Santus. 2021. BERT Prescriptions to Avoid Unwanted Headaches: A Comparison of Transformer Architectures for Adverse Drug Event Detection. In *Proceedings of EACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Mingyu Wan, Rong Xiang, Emmanuele Chersoni, Natalia Klyueva, Kathleen Ahrens, Bin Miao, David Clive Broadstock, Jian Kang, Hing Wah Yung, and Chu-Ren Huang. 2019. Sentence Boundary Detection of Financial Data with Domain Knowledge Enhancement and Cross-lingual Training. In *Proceedings of The First Workshop on Financial Technology and Natural Language Processing: The FinSBD Shared Task*.

Rong Xiang, Emmanuele Chersoni, Luca Iacoponi, and Enrico Santus. 2020. The CogALex Shared Task on Monolingual and Multilingual Identification of Semantic Relations. In *Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon*.

Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets. In *Proceedings of COLING*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. FinBERT: A Pretrained Language Model for Financial Communications. *arXiv preprint arXiv:2006.08097*.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2020. Revisiting Few-sample BERT Fine-tuning. In *Proceedings of ICLR*.

Siyan Zheng, Anneliese Lu, and Claire Cardie. 2020. SUMSUM@ FNS-2020 Shared Task. In *Proceedings of the Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27.

A Appendix

Figure 1 shows the pie charts illustrating the distribution of classes for all the benchmark datasets.

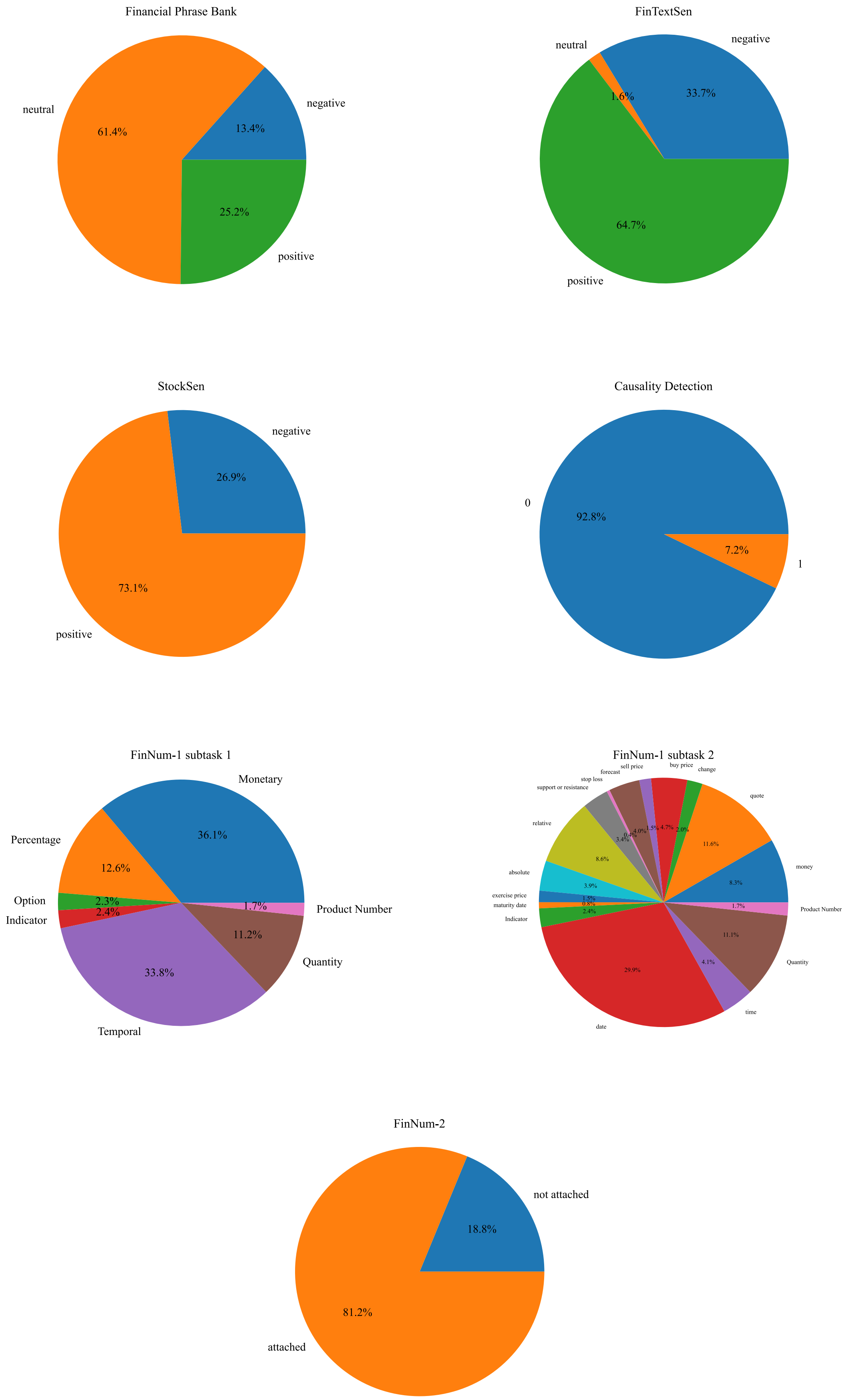

Figure 1: Class distribution for each of the evaluation datasets.