FiNER-ORD: Financial Named Entity Recognition Open Research Dataset

Agam Shah , Abhinav Gullapalli , Ruchit Vithani , Michael Galarnyk , Sudheer Chava
Georgia Institute of Technology

Abstract

The development of annotated datasets has contributed to the power of deep learning and natural language processing (NLP). Most datasets created for named entity recognition (NER) are not domain-specific. The finance domain presents specific challenges to the NER task and a domain-specific dataset would help push the boundaries of finance research. In our work, we develop the first high-quality English Financial NER Open Research Dataset (FiNER-ORD). We benchmark multiple pre-trained language models (PLMs) and large-language models (LLMs) on FiNER-ORD. We believe our proposed FiNER-ORD dataset will open future opportunities to use FiNER-ORD as a primary benchmark for financial domain-specific NER and NLP tasks. Our dataset, models, and code will be made publicly available on GitHub and Hugging Face under CC BY-NC 4.0 license.

1 Introduction

The growth of technology and the web over the last several decades has led to a rapid increase in the generation of text data, especially in the financial domain. The abundance of financial texts, primarily in the form of news and regulatory writings, presents a valuable resource for analysts, researchers, and even individual investors to extract relevant information. However, processes for information retrieval must be efficient to make data-driven decisions in a timely manner. With the given abundance of text data, manually extracting relevant information is impossible and unsustainable for large-scale text datasets used for rapid downstream tasks. However, natural language processing (NLP) techniques can help with automating the information retrieval process. Named entity recognition (NER) is one such NLP technique that serves as an important first step to identify named entities, such as persons, organizations, and locations, and efficiently use available text data to ultimately drive downstream tasks and decisions.

Correspondence to Agam Shah {ashah482@gatech.edu}

While numerous studies have constructed annotated datasets (Sang and De Meulder, 2003; Derczynski et al., 2017a; Weischedel et al., 2013) and developed NER models (Li et al., 2019; Yamada et al., 2020; Wang et al., 2021a,b) for generic texts, the financial domain presents unique challenges that require domain-specific approaches and expertise. The current annotated dataset based on Credit Risk Agreements (Alvarado et al., 2015) for financial NER has significant shortcomings, limiting the use of deep learning models for financial NER. We elaborate on these shortcomings in Section 3.2.

In addressing these challenges within the realm of resources and evaluation, our work adds value to the field by establishing the largest Financial Named Entity Recognition Open Research Dataset (FiNER-ORD). Our objective in this context is not to introduce a novel state-of-the-art (SOTA) model architecture. Instead, we undertake a benchmarking exercise, evaluating pre-trained language models (PLMs) and zero-shot large-language models (LLMs) using the FiNER-ORD dataset.

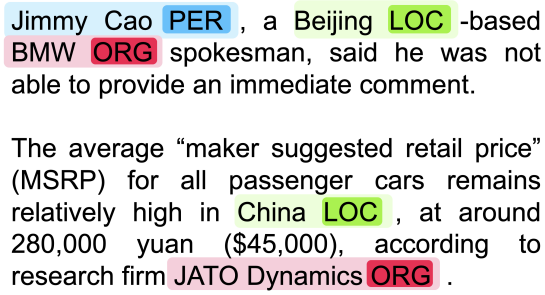

Figure 1: Representative example of annotation in FiNER-ORD.

2 Related Work

Named Entity Recognition Datasets Commonly used general NER datasets include the

CoNLL-2003 (Sang and De Meulder, 2003), OntoNotes5.0 (Weischedel et al., 2013), WNUT 2017 (Derczynski et al., 2017b), and FewNERD (Ding et al., 2021). FiNER-139 from Loukas et al. (2022) is an entity recognition dataset for numerical financial data. Like our work, they also don't use numerical entities because they can be easily extracted using regular expressions and are the easiest entity types to recognize. There has also been work in recognizing entities in invoices, business forms, and emails (Francis et al., 2019). CRA NER dataset from Alvarado et al. (2015) was the first attempt to create a financial NER dataset similar to FiNER-ORD, but in our assessment presented in Section 3.2, it is skewed and thus limited in usefulness.

Applications for Information Retrieval in Finance Sentiment analysis of news aids event ranking based on market factors, aiding price prediction (Feng et al., 2021). Analyzing tweets and headlines automates trading, price movement, and risk forecasts (Sawhney et al., 2021). Extracting claims from analysts' reports improves volatility forecasting on release and earning dates (Shah et al., 2024). Real-time web tools extract operating segments, aiding performance analysis (Ma et al., 2020). Transformer models fine-tuned for price-change data extraction measure inflation exposure (Chava et al., 2022). Financial knowledge graphs from news data enhance algorithmic trading (Cheng et al., 2020).

3 Dataset

3.1 FiNER-ORD

FiNER-ORD consists of a manually annotated dataset of financial news articles (in English) collected from webz.io[1]. In total, there are 47,851 news articles available for download in this dataset at the time of writing this paper.

Sampling and Manual Annotation For the manual annotation of named entities in financial news, we randomly sampled 220 documents from the entire set of news articles. We observed that some articles were empty in our sample, so after filtering the empty documents, we were left with a total of 201 articles. We use the open-source Doccano[2] annotation tool (see Appendix E for more details) to ingest the raw dataset and manually label person (PER), location (LOC), and organization (ORG) entities. Figure 1 shows an example of manually annotated named entities in FiNER-ORD. Labeling was performed independently by two different annotators, following the predefined annotation guide presented in Appendix D, to reduce potential labeling bias. Annotator labeling was compared and validated for consistency. We provide details about annotator background in Appendix A, annotator agreement in Appendix B, and post-processing in Appendix C.

FiNER-ORD	Train	Validation	Test
Articles	135	24	42
Sentences	3,262	402	1,075
Tokens	80,531	10,233	25,957
PER	821	138	284
LOC	966	193	300
ORG	2,026	274	544

Table 1: Number of articles, sentences, tokens, and entities (person, location, organization) in the train, validation, and test splits of FiNER-ORD.

Dataset Statistics Following the manual annotation and post-processing procedures, each token is labeled as one of 4 broad entity types: PER, LOC, ORG, and O. As discussed in Appendix C each of the PER, LOC, and ORG classes are further segmented with the suffixes _B (denoting beginning token of a span) and _I (denoting intermediate token of a multi-token span), respectively. As discussed in section 3.1, we manually annotate the train, validation, and test splits of the 201 articles in FiNER-ORD. The descriptive statistics on the resulting FiNER-ORD are available in Table 1.

3.2 Comparison of Datasets

Credit Risk Agreements (CRA) Table 2 compares our proposed dataset (FiNER-ORD) with the CRA NER dataset (Alvarado et al., 2015). The dataset attempts to provide a domain-specific dataset in CoNLL format using manual annotation on eight English documents from the U.S. Security and Exchange Commission (SEC) filings (Bird et al., 2009). Unfortunately, the annotation methodology for the CRA NER dataset "automatically tagged all instances of the tokens *lender* and *borrower* as being of entity type PER" (Alvarado et al., 2015). This approach is problematic because of the resulting skewed distribution of entity types in the dataset, leading to confounded results. Our analysis of the CRA NER dataset showed that in FIN3 (CRA

[1] https://webz.io/free-datasets/financial-news-articles/

[2] https://github.com/doccano/doccano

test data split), instances of the tokens *lender* and *borrower* represented 83.05% of all PER tokens and 44.95% of all tokens labeled as PER, ORG, MISC, or LOC. Similarly, in FIN5 (CRA train data split), instances of the tokens *lender* and *borrower* represented 90.04% of all PER tokens and 46.08% of all tokens labeled as PER, ORG, MISC, or LOC.

Thus, we believe the CRA dataset is not a high-quality benchmark for specialized NLP tasks in the financial domain, motivating us to create a new high-quality financial domain-specific NER dataset. The dataset comparison statistics are available in Table 2. We also perform a transfer learning study presented in Section 5.1 in order to further emphasize the importance of FiNER-ORD in comparison to the CRA NER dataset.

CoNLL-2003 The CoNLL NER dataset (Sang and De Meulder, 2003) was created with manual annotation on Reuters generic news stories which were pre-tokenized and part-of-speech (POS) tagged by the memory-based MBT tagger (Daelemans et al., 2002). In general, financial texts differ from general texts and contain a higher ratio of organization tokens and entities compared to person and location tokens and entities. This can be important for financial news network applications that need to capture the interdependence of companies and the complexity of modern financial markets. In the FiNER-ORD dataset, the ratio of organizations (ORG) to location (LOC) to person (PER) entities is 2.29:1.17:1. This distribution contrasts with the CRA and CoNLL datasets, which exhibit ORG:LOC:PER ratios of 0.31:0.22:1 and 0.93:1.06:1, respectively. We also analyze the average length of each entity-type in terms of the number of tokens per entity. In the order of FiNER-ORD, CRA, and CoNLL, we see the average number of tokens for a PER entity to be approximately 1.66, 1.06, and 1.69, respectively. Similarly, for LOC entities we find an average length of approximately 1.34, 2.09, and 1.50, respectively. For ORG entities, we find an average length of 1.72, 1.69, 1.18, respectively. These findings show that for PER entities, the average length is similar in FiNER-ORD and CoNLL. However, the skewed distribution of PER entities with the problematic labeling of *lender* and *borrower* as PER in CRA leads to an average length value close to 1 for PER entities in CRA, underscoring the finance-specific utility of FiNER-ORD over CRA. Even with LOC entities, the average length values are similar in FiNER-ORD and CoNLL. However, the finance-specific usefulness of FiNER-ORD over CoNLL is highlighted by the average length values for ORG entities. We see similar average lengths for ORG entities in FiNER-ORD and CRA, both of which are financial NER datasets, but the average length of ORG entities in CoNLL is much smaller. Additionally, the comparison between the ratios of ORG_B and ORG_I tokens for FiNER-ORD (1.4:1) and CoNLL (5.6:1) suggests that in financial texts, organization entities are more likely to span over multiple tokens. Having a higher percentage of useful ORG entities and tokens can also enhance applications like measuring correlations in financial news networks and market movements (Wan et al., 2021). This highlights the importance of creating high-quality datasets specifically for the financial domain.

Dataset	FiNER-ORD	CRA	CoNLL
Articles	201	8	1,393
Sentences	4,739	1,473	22,137
Tokens	116,721	54,262	301,418
PER	1,243	962	10,059
LOC	1,459	208	10,645
ORG	2,844	295	9,323

Table 2: Comparison of our FiNER-ORD with Credit Risk Agreements and CoNLL-2003 English NER datasets in terms of the number of articles, sentences, tokens, and entities (person, location, organization).

4 Models

4.1 PLMs

We benchmark FiNER-ORD with several base and large transformer-based models. For the base model category, we use BERT (Devlin et al., 2018), FinBERT (Yang et al., 2020), and RoBERTa (Liu et al., 2019). For the large model category, we use BERT-large (Devlin et al., 2018) and RoBERTa-large (Liu et al., 2019). We do not pre-train these models before fine-tuning them. Further details about fine-tuning is provided in Appendix F.

4.2 LLMs

To benchmark the performance of current SOTA generative LLMs, we measure the zero-shot performance on three train validation splits for the "gpt-4-0613" and "gpt-3.5-turbo-0613" models[3] with 0.0 temperature and 1000 max tokens for the output. All API calls were made on October 11, 2023 or

[3] https://platform.openai.com/docs/guides/gpt

Model	Train Split	PER	LOC	ORG	Weighted Average
Panel A: FineTuning with PLMs					
BERT-base-cased	FiNER-ORD	0.8811 (0.0192)	0.6820 (0.0239)	0.6013 (0.0490)	0.6931 (0.0327)
FinBERT-base-cased	FiNER-ORD	0.7456 (0.0254)	0.6836 (0.0238)	0.6002 (0.0288)	0.6589 (0.0231)
RoBERTa-base	FiNER-ORD	0.9050 0.0076)	0.7154 (0.0608)	0.6304 (0.0878)	0.7220 (0.0585)
BERT-large-cased	FiNER-ORD	0.8954 (0.0090)	0.7289 (0.0467)	0.6272 (0.0145)	0.7216 (0.0039)
RoBERTa-large	FiNER-ORD	**0.9263** (0.0025)	**0.7717** (0.0152)	**0.6769** (0.0130)	**0.7648** (0.0057)
Panel B: Zero-Shot with Generative LLMs					
ChatGPT-3.5-Turbo	Zero-Shot	0.6757 (0.0037)	0.5752 (0.0059)	0.4049 (0.0015)	0.5178 (0.0009)
GPT-4	Zero-Shot	0.7425 (0.0053)	0.7232 (0.0067)	0.5754 (0.0074)	0.6563 (0.0038)
Panel C: Transfer Learning Ablation					
RoBERTa-large	CRA	0.5918 (0.0868)	0.0145 (0.0020)	0.0411 (0.0174)	0.1730 (0.0262)
RoBERTa-large	CoNLL	0.8707 (0.0171)	0.7640 (0.0278)	0.5668 (0.0678)	0.6954 (0.0278)

Table 3: Performance comparison of various models tested on the FiNER-ORD test split. Panel A contains results for various models fine-tuned on the FiNER-ORD train sample. Panel B contains results for zero-shot GPT models. Panel C contains results for transfer learning experiments where the RoBERTa-large model is fine-tuned on the Credit Risk Agreements (CRA) NER dataset and CoNLL-2003 dataset. All values are weighted F1 scores. An average of 3 seeds was used for all models. The standard deviation of the F1 scores is reported in parentheses.

October 12, 2023. The zero-shot prompt can be found in Appendix G.

5 Results and Analysis

In this section, we evaluate and benchmark different NLP models on the NER task with FiNER-ORD. For all models, we consistently use the same train, validation, and test splits in FiNER-ORD at the sentence-level.

For each PLM, we use three seeds (5768, 78516, 944601) for the three runs of each model. We run each LLM three times with the same settings discussed in Section 4.2. We report average weighted F1 scores for the best hyper-parameter configuration of each model in Table 3.

Although zero-shot GPT-4 outperforms zero-shot ChatGPT-3.5-Turbo overall, fine-tuned PLMs outperform both zero-shot LLMs across all entity label categories. This finding aligns with the survey by Pikuliak (2023), which finds that zero-shot ChatGPT fails to outperform fine-tuned models on more than 77% of NLP tasks. RoBERTa-base achieves best weighted F1 score overall and for all entity label categories when compared to all models. A future ablation study of these language models would provide key insights such as why RoBERTa-base, a model more generally trained with the masked language modeling (MLM) objective, outperforms (when considering our financial NER evaluation) FinBERT-base, a model trained on financial sentiment classification tasks and widely used for various financial domain-specific NLP tasks. Applying FiNER-ORD as a financial-domain NER benchmark dataset for such studies can further improve financial domain-specific language models.

5.1 Transfer Learning Ablation

This study evaluates the transfer learning capabilities of the best-performing model, RoBERTa-large, originally trained and tested on the FiNER-ORD dataset. Specifically, we investigate its performance when fine-tuned on the CRA NER dataset and CoNLL dataset, followed by testing on the test split of the FiNER-ORD dataset. The outcomes of these transfer learning experiments are detailed in Panel C of Table 3. We also evaluate an existing model trained on the CONNL 2003 dataset and test on FiNER-ORD in Table I.2. These results underscore the significance of the FiNER-ORD dataset in enhancing model performance relative to the existing CRA and CoNLL NER datasets.

6 Conclusion

We present the first high-quality, manually annotated financial NER dataset, FiNER-ORD, which was generated from open-source financial news articles. We demonstrate the importance of our FiNER-ORD dataset compared to the existing CoNLL-2003 and financial CRA NER datasets. To evaluate the proposed dataset, we benchmark various configurations of PLMs and LLMs on FiNER-ORD. The performance analysis shows that RoBERTa-base outperforms all tested models (including FinBERT-base) overall, and both GPT-4 and Chat-GPT-3.5-Turbo underperform the tested PLMs over all entity label categories. Furthermore, our proposed financial domain-specific dataset and performance analysis together present opportunities to further explore improvements for financial domain-specific language models.

Ethics Statement

All language models used, under their respective license categories, are publicly available for our experimental purposes. With regards to ethical considerations for the environmental impact of our experiments, we only fine-tune the benchmarked PLMs because pre-training PLMs are known to have a large carbon footprint. We acknowledge that our dataset is constructed using English news articles and thus biased in terms of language representation and inclusion. The news articles do not contain offensive or discriminatory content. The dataset license and copyright exceptions are stated in Appendix H.

Limitations

We acknowledge a geographic bias in the dataset as FiNER-ORD only includes English financial news articles. In the future, other forms and languages of financial texts could be manually annotated to expand FiNER-ORD. Additionally, more label classes such as 'product', 'miscellaneous', and forms of relationships could be annotated for future downstream applications such as knowledge graph creation. We do not present a new SOTA model architecture for financial NER and we do not include or discuss finance domain-specific LLMs like BloombergGPT (Wu et al., 2023) in our work as we have no way of accessing it. Our primary focus of this work is to present the first high-quality financial NER open research dataset and benchmark our FiNER-ORD with multiple PLMs and LLMs to evaluate the performance of these models on the finance domain-specific NER task.

Form 10-K Filings When considering data sources for financial NER, we found that SEC Form 10-K filings, filed annually by public companies in the United States, have certain limitations. The 10-K provides details about a company's business, risks, and operating and financial results for each fiscal year. For the NER task, however, we observe there is a noticeable bias towards organization entity tokens associated with the company filing the 10-K. As a result, there is a reduced diversity of named entities, particularly concerning other organizations. Furthermore, the average length of 10-K filings has been growing significantly every year with at least 60,000 words between 2022 and 2023, a significant increase from around 30,000 words in 2000 and approximately 42,000 words in 2013[4]. In contrast, financial news articles are shorter and can be used to better capture the interdependence of companies and the complexity of modern financial markets. Given this, we decided that constructing a high-quality, manually annotated dataset would be more feasible and information-diverse with the open-source news articles we annotate in FiNER-ORD as opposed to 10-K filings.

References

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90.

Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural language processing with python — analyzing text with the natural language toolkit.

Sudheer Chava, Wendi Du, Agam Shah, and Linghang Zeng. 2022. Measuring firm-level inflation exposure: A deep learning approach. *Available at SSRN 4228332*.

Dawei Cheng, Fangzhou Yang, Xiaoyang Wang, Ying Zhang, and Liqing Zhang. 2020. Knowledge graph-based event embedding framework for financial quantitative investments. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2221–2230.

Walter Daelemans, Jakub Zavrel, Antal van den Bosch, and Ko van der Sloot. 2002. MBT: Memory-based tagger version 1.0 reference guide. *ILK Technical Report ILK-0209, University of Tilburg, The Netherlands*.

Leon Derczynski, Eric Nichols, Marieke Van Erp, and Nut Limsopatham. 2017a. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017b. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[4] https://www.wsj.com/articles/the-109-894-word-annual-report-1433203762

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213, Online. Association for Computational Linguistics.

Fuli Feng, Moxin Li, Cheng Luo, Ritchie Ng, and Tat-Seng Chua. 2021. Hybrid learning to rank for financial event ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 233–243.

Sumam Francis, Jordy Van Landeghem, and Marie-Francine Moens. 2019. Transfer learning for named entity recognition in financial and biomedical documents. *Information*, 10(8).

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2019. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431, Dublin, Ireland. Association for Computational Linguistics.

Zhiqiang Ma, Steven Pomerville, Mingyang Di, and Armineh Nourbakhsh. 2020. Spot: A tool for identifying operating segments in financial tables. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2157–2160.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Matúš Pikuliak. 2023. Chatgpt survey: Performance on nlp datasets. https://www.opensamizdat.com/posts/chatgpt_survey.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Ramit Sawhney, Shivam Agarwal, Megh Thakkar, Arnav Wadhwa, and Rajiv Ratn Shah. 2021. Hyperbolic online time stream modeling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1682–1686.

Agam Shah, Arnav Hiray, Pratvi Shah, Arkaprabha Banerjee, Anushka Singh, Dheeraj Eidnani, Bhaskar Chaudhury, and Sudheer Chava. 2024. Numerical claim detection in finance: A new financial dataset, weak-supervision model, and market analysis.

Xingchen Wan, Jie Yang, Slavi Marinov, Jan-Peter Calliess, Stefan Zohren, and Xiaowen Dong. 2021. Sentiment correlation in financial news networks and associated market movements. *Scientific Reports*, 11:3062.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021a. Automated Concatenation of Embeddings for Structured Prediction. In *the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (**ACL-IJCNLP 2021**)*. Association for Computational Linguistics.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021b. Improving named entity recognition by external context retrieving and cooperative learning. *arXiv preprint arXiv:2105.03654*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.1564*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *CoRR*, abs/2006.08097.

A Annotator Background

Labeling was performed independently by two different annotators. Annotator 1 is a Doctoral Researcher from India working with NLP and Computational Finance. Annotator 2 is a Masters student from the United States pursuing a Computer Science degree with Machine Learning specialization. Both annotators were male researchers and read English financial news articles daily from several outlets and sources. Additionally, both annotators are authors, so the annotators were not hired or paid for annotations.

B Annotator Agreement

Both annotators had a manual annotation agreement of approximately 96.85% of the 5546 entities across the PER, LOC, ORG entity classes and train, validation, test splits of FiNER-ORD. The annotation guide was referenced to resolve disagreements between the manual labeling of both annotators. The annotation guide was developed iteratively during the annotation process. Online resources were consulted if the annotation guide did not address a specific disagreement, and the annotation guide was updated accordingly afterwards.

C Post-Processing

To correct potential errors in manual annotations, we run a custom post-processing script that performs the following four tasks: (1) remove trailing spaces from annotated entities, (2) extend token-level borders to non-space characters to change an erroneous span to the correct span, (3) clean entity suffixes with techniques such as removing an apostrophe followed by the letter s (*'s*) from entity suffixes, (4) tokenize text with Stanza (Qi et al., 2020) and add positional information for labeled entities by splitting multi-token spans into separate tokens, assigning _B as the label suffix for the first separated token in the multi-token span, and assigning _I as the label suffix for the remaining separated tokens in the multi-token span. We note that after the post-processing script is run, all tokens which are not annotated with one of PER_B, PER_I, LOC_B, LOC_I, ORG_B, ORG_I are assigned the label O, denoting "other" type of token not belonging to the person, location, organization classes.

D Annotation Guide

The manual annotation process to create FiNER-ORD consisted of ingesting the financial news articles in Doccano. Each news article is available in the form of a JSON document with various metadata information including the source of the article, publication date, author of the article, and the title of the article. Entities of the type *person* (PER), *organization* (ORG), and *location* (LOC) were identified according to the rules described below. Some well-known names for these entities were obvious while others were confirmed by researching the names to identify the correct entity type.

D.1 Person Entities

PER entities were identified by their first name and/or last name. In the examples below, bold spans represent a single person entity. In the case where a person was identified by their first and last name, the entire name was labeled as PER and the post-processing script tagged the first name as PER_B and the last name as PER_I. Words like *President*, *Ms*, and *CEO* were not labeled as part of the PER entity but help indicate a PER entity. In a context indicating possession with *'s*, the name until the *'s* was tagged.

- President **Obama**
- CEO **Phyllis Wakiaga**
- Ms **Wakiaga**
- **Bill Clinton**'s

D.2 Location Entities

LOC entities primarily consisted of names of continents, countries, states, cities, and addresses. In the examples below, bolded spans represent spans comprising LOC entities. Commas in addresses are not included tagged LOC entities. In such cases for tagging addresses, each complete span delimited by a comma was tagged as a LOC entity. In

a context indicating possession with *'s*, the name until the *'s* was tagged. In the case where a location was identified by multiple tokens delimited by a space, the entire name was labeled as LOC and the post-processing script tagged the first name as LOC_B and the last name as LOC_I. Words such as "Kenyan" were treated as adjectives and thus not labeled as a LOC entity. When discussing a lawmaker's political and location affiliation, examples such as *R-Texas* denoting "Republican from Texas" are encountered, in which only the location name such as *Texas* is tagged.

- **Asia**
- **US**
- **India**
- **United States**
- **Beijing**
- **New York**
- **Redwood City**, **California**
- **Kenya**'s
- **Mombasa Road**
- R-**Texas**

D.3 Organization Entities

ORG entities consist of examples such as company names, news agencies, government entities, and abbreviations such as stock exchange names and company stock tickers. Punctuation marks such as hyphens are included when tagging an ORG entity. As designated, *.com* is included in the identified company's name. In a context indicating possession with *'s*, the name until the *'s* was tagged.

- **Wal-Mart**
- **China Resources SZITIC Trust Co Ltd**
- **The Wall Street Journal**
- **Atlanta Federal Reserve**
- **Morgan Stanley**'s
- **Delta Air Lines**
- **DAL**
- **NYSE**
- **Amazon.com**

D.4 Most Common Entities

Table 4 shows the most common entities within FiNER-ORD. Since the dataset is comprised of financial news articles from July to October 2015, the dataset has a bias towards the news of that time period. An advantage of FiNER-ORD is that it has a unique heterogeneity due to it being composed of English language financial articles published by institutions from around the world unlike the CRA dataset which are from United States Securities and Exchange Commission documents. The most common LOC and ORG entities reflect the global news article source.

D.5 Annotation Edge Cases

There were a few edge cases in the annotation process. For example, labeling entities when location (LOC) is part of the organization (ORG) entity is a common problem in finance. The phrase "Google India" has "India" which is a location, but it is labeled as an organization in our framework. This is because our process does not permit overlapping entity labels. Another such example is "New York Stock Exchange" which we annotated entirely as an ORG entity, despite models often predicting "New York" as LOC. Therefore, the correct labels are ORG_B for "New" and ORG_I for "York," whereas models might label them as LOC_B and LOC_I, respectively. Commonly used words like "the" which are often tagged as O (Other) may sometimes be present in the name of an organization, such as "The Wall Street Journal". In such cases, we have tagged "The" to be included as part of an ORG entity. In the specific example for "The Wall Street Journal", as shown in Appendix D, we have tagged "The" as ORG_B and the remaining tokens "Wall", "Street", "Journal" as ORG_I tokens. For entities which may represent both an organization and a product, such as "Google", we have only tagged such entities as ORG because our dataset currently does not provide manual annotations for product entities.

E Doccano Annotation Tool

All manual annotation of FiNER-ORD was completed using the open-source Doccano[5] annotation tool. Figure 2 demonstrates the use of Doccano to manually annotate FiNER-ORD. The output from Doccano contains span-level label information. This information is in the form of a list of lists

[5] https://github.com/doccano/doccano

All Entities	PER	LOC	ORG
China-155	Obama-26	China-154	GM-52
U.S.-69	Abbott-19	U.S.-69	Reuters-47
GM-52	Clinton-18	Greece-49	Facebook-35
Greece-49	Bush-14	US-46	Nikkei-34
Reuters-47	Turnbull-14	UK-37	Fed-32
US-46	Varoufakis-11	Australia-32	Apple-31
UK-37	Jaitley-9	London-28	Ford-24
Facebook-35	Shorten-9	Japan-24	Stratasys-23
Nikkei-34	McPhail-9	New York-23	LSE-21
Australia-32	Tony Abbott-8	Europe-22	House-20
Fed-32	Malcolm Turnbull-8	Sydney-21	UK Markets-19
Apple-31	Andrew-8	Taiwan-19	Unilever-19
London-28	Bishop-8	Bahrain-18	Motley Fool-18
Obama-26	John - Erik Koslosky-8	Kenya-18	SEC-18
Japan-24	Glatt-7	India-17	FT-18

Table 4: Most common entities in FiNER-ORD.

Model	learning rate	batch size
BERT-base-cased	1e-5	16
FinBERT-base-cased	1e-5	8
RoBERTa-base	1e-5	8
BERT-large-cased	1e-5	8
RoBERTa-large	1e-5	8

Table 5: Best hyper-parameter configuration for each PLM benchmarked on FiNER-ORD.

containing information on the start character, end character, and label of each entity annotated by the manual annotator.

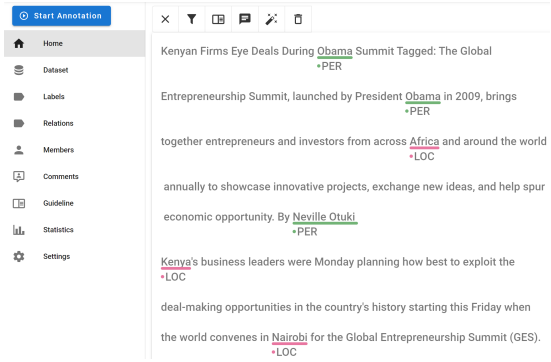

Figure 2: Screenshot of an article in FiNER-ORD manually annotated with the open-source Doccano annotation tool.

F Fine-tuning PLM Details

No pre-training is conducted on the models before proceeding with the fine-tuning process. To determine the most suitable hyper-parameters for each model, we performed a grid search using three different learning rates (1e-4, 1e-5, 1e-6) and three different batch sizes (32, 16, 8). We use a maximum of 100 epochs for training with early stopping criteria. If the validation F1 score doesn't improve by more than or equal to 1e-2 in the next 7 epochs then we use the best model stored earlier as the final 6 fine-tuned model. We report the best hyper-parameter configuration for each PLM in Table 5. All of our experiments are carried out using PyTorch (Paszke et al., 2019) on an NVIDIA RTX A6000 GPU. Each model is initialized with the pre-trained version available in the Hugging Face Transformers library (Wolf et al., 2020).

G Zero-shot Prompt

"Discard all the previous instructions. Behave like you are an expert named entity identifier. Below a sentence is tokenized and each line contains a word token from the sentence. Identify "Person", "Location", and "Organisation" from them and label them. If the entity is multi token use post-fix _B for the first label and _I for the remaining token labels for that particular entity. The start of the separate entity should always use _B post-fix for the label. If the token doesn't fit in any of those three categories or is not a named entity label it 'Other'. Do not combine words yourself. Use a colon to separate token and label. So the format should be token:label. {sentence}".

H Copyright Exceptions

The dataset will be made publicly available on Hugging Face under the non commercial CC BY-NC 4.0 license. The individual articles comprising the dataset can be considered exempt from copyright for non-commercial research. Exceptions to copyright in the United Kingdom are regularly updated at gov.uk[6]. The EU Directive on Copyright in the

[6] https://www.gov.uk/guidance/exceptions-to-copyright

Train Split	Test Split	PER	LOC	ORG	Weighted Avg.
FiNER	FiNER	**0.9263 (0.0025)**	0.7717 (0.0152)	**0.6769 (0.0130)**	**0.7648 (0.0057)**
CRA	FiNER	0.5918 (0.0868)	0.0145 (0.0020)	0.0411 (0.0174)	0.1730 (0.0262)
CoNLL	FiNER	0.8707 (0.0171)	0.7640 (0.0278)	0.5668 (0.0678)	0.6954 (0.0278)
FiNER+CoNLL	FiNER	0.9107 (0.0185)	**0.7992 (0.0071)**	0.6441 (0.0183)	07522 (0.0141)
FiNER	CoNLL	0.9164 (0.0299)	0.7099 (0.0218)	0.5622 (0.0439)	0.7278 (0.0317)
CRA	CoNLL	0.6498 (0.0562)	0.0492 (0.0173)	0.2078 (0.0685)	0.2988 (0.0316)
CoNLL	CoNLL	0.9553 (0.0080)	0.8974 (0.0214)	**0.8038 (0.0180)**	0.8849 (0.0146)
FiNER+CoNLL	CoNLL	**0.9559 (0.0047)**	**0.9018 (0.0102)**	0.8001 (0.0965)	**0.8866 (0.0317)**

Table 6: Cross-dataset and combined dataset performance analysis of FiNER, CRA, and CoNLL

Digital Single Market[7] provides exceptions for reproductions made by research organizations.

I Comparison with CoNLL Dataset

I.1 Combined Dataset Performance

In order to test whether FiNER-ORD and CoNLL can be used together to achieve better NER performance, we combine the training data from both datasets. We test separately on the FiNER-ORD and CoNLL test sets. The Weighted Avg. results shown in Table 6 suggest that when testing on CoNLL, FiNER-ORD can be used to complement CoNLL.

I.2 Training on CoNLL and Testing on FiNER-ORD

Section 3.2 highlighted how financial texts differ from general texts. In particular, financial texts tend to contain a higher ratio of organization tokens and entities compared to person and location tokens and entities. To evaluate how a model trained on the non finance specific CoNLL dataset would do on the FiNER-ORD test split, we evaluate an existing model trained on the CONNL 2003 dataset[8]. This model only performs better on LOC entities than a similar model trained on the FiNER-ORD train split.

[7] https://eur-lex.europa.eu/eli/dir/2019/790/oj

[8] https://huggingface.co/tner/roberta-large-conll2003

Model	Test Split	PER	LOC	ORG	Weighted Avg.
tner/roberta-large-conll2003	FiNER	0.9062	0.8788	0.8004	0.8474
roberta-large-finer	FiNER	0.9384	0.8520	0.8352	0.8637

Table 7: Transfer learning ablation on CoNLL using tner framework