Financial Entity Recognition with FinBERT

Sameer Karim
sameerkarim@berkeley.edu

Abstract

Automated extraction of financial entities from financial disclosures is essential for enhancing decision-making and maintaining regulatory compliance. This study investigates the application of FinBERT, a transformer-based NLP model specifically fine-tuned for financial text, in performing Named Entity Recognition (NER). Employing detailed preprocessing strategies and rigorous hyperparameter tuning, the model achieved a marked improvement, yielding an F1-score of 58.41%. This improvement from an initial baseline of 48% underscores the efficacy, while including room for improvement, of domain-specific transformer models in extracting financial entities. Furthermore, the study provides comprehensive insights into the methodological choices and highlights potential avenues for future research to address current challenges.

Introduction & Background

Transformer-based architectures like BERT have significantly improved NLP tasks over traditional methods (HMMs, CRFs). However, general BERT models, while strong, often struggle with domain-specific applications due to training on general datasets. FinBERT, trained specifically on financial texts (Araci, 2019), has already demonstrated superior performance in finance-related sentiment analysis. Its potential in NER tasks, however, has remained largely unexplored.

The automated extraction of named entities from financial documents is critically important due to the vast amounts of data produced and utilized within financial markets. Financial disclosures, regulatory filings, and corporate financial statements contain extensive specialized terminology and structures, making entity extraction both challenging and necessary. Traditionally, financial entities are identified manually by domain experts, a process that is labor-intensive, costly, and susceptible to errors.

Objective & Data

The goal of this research is to address this gap by rigorously evaluating FinBERT's capability in financial NER tasks, extending the current understanding of domain-specific transformer applications and clearly distinguishing itself from recent models like FinQANet (Chen et al., 2022) and Finner (Yang et al., 2021) through extensive preprocessing and hyperparameter optimization. We will explore this idea by developing an automated system for identifying and classifying financial entities based on 139 unique XBRL tags from financial disclosures (derivative of the FiNER-139 dataset). Data acquisition from a Financial NER dataset from Hugging Face, consisting of approximately 900,000 financial disclosures annotated with 139 XBRL tags, was utilized *(https://huggingface.co/datasets/Josephgflowers/Financial-NER-NLP).* This new dataset transforms the original structured data from

FiNER-139 into natural language prompts suitable for training language models. The dataset is designed to enhance models' abilities in tasks such as named entity recognition (NER), summarization, and information extraction in the financial domain.

Performance metrics like precision, recall, and F1-score, were selected due to their balance between false positives and false negatives, which are critical in financial contexts. An initial baseline was established using a FinBERT fine-tuning approach, resulting in an F1-score of approximately 48%. The goal was to surpass this baseline significantly through preprocessing and hyperparameter optimization strategies.

Methodology & Model Tuning

Due to computational constraints, the dataset was strategically downsampled to 50,000 training samples and 10,000 test samples, randomly selected to maintain representativeness.

Preprocessing steps were crucial for achieving high-quality results. Unicode normalization standardized text encoding across the dataset. Tokenization combined whitespace-based splitting for compatibility with FinBERT's tokenizer and regular expression-based methods to accurately capture financial terminologies and numerical data. Labels were extracted from JSON-formatted annotations, ensuring precise alignment between tokens and their corresponding labels. These preprocessing changes significantly improved recall by capturing entities previously missed due to tokenization mismatches, resulting in a recall increase from 54% to 60.82%.

Mixed precision training (fp16) was employed to leverage GPU efficiency, significantly accelerating the training process without sacrificing model accuracy. Warmup steps mitigated potential abrupt learning rate changes at training onset, reducing initial training instability and enhancing model convergence.

Results

The refined methodology significantly improved model performance over the baseline. Precision increased substantially from approximately 42% to 56.17%, primarily due to stable and gradual learning enabled by the adjusted hyperparameters. Recall showed remarkable improvement, rising from 54% to 60.82%, driven largely by more effective preprocessing strategies that minimized tokenization errors and label misalignments. Consequently, the overall F1-score improved notably, increasing from the baseline of 48% to 58.41%. Analysis of training epochs revealed rapid early improvements followed by diminishing returns, suggesting that further gains may require innovative approaches beyond conventional hyperparameter tuning.

	Model	Test Precision	Test Recall	Test F1	Test Validation Loss
1	Baseline	0.500502	0.473435	0.486592	0.007244
2	Second Model	0.492417	0.462049	0.47675	0.007326
3	Final Model	0.570906	0.59203	0.581276	0.009241

Fig. 1 Results

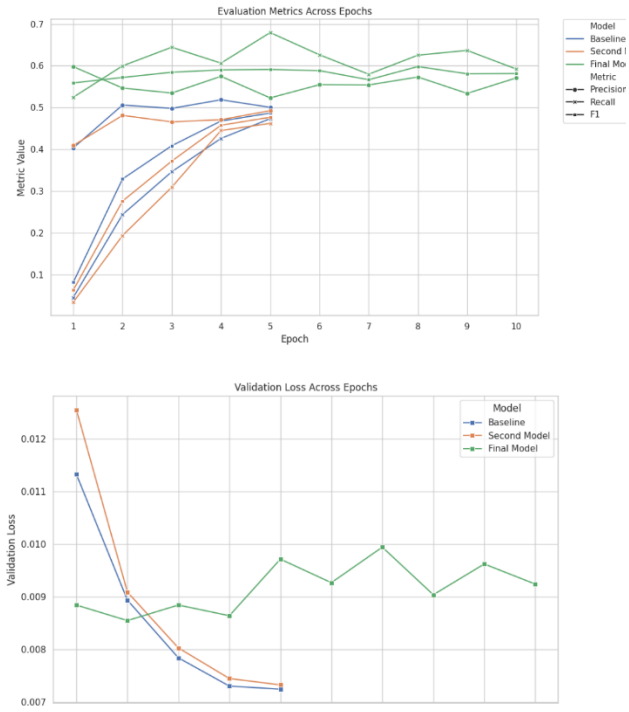

Fig. 2 Metric Improvement

Discussion

Error Analysis

Despite improved performance, structured error analysis revealed several recurring misclassification patterns at the token and span level. These failure cases shed light on why precision and recall plateaued and identify where the FinBERT-based model struggles with financial entity recognition. The main error types and their impact on precision, recall, and F1-score follows as:

Fragmented or Partial Entity Recognition: The model often failed to capture full multi-token entities, especially compound financial terms. In many cases, only a part of a compound entity was recognized, leaving the rest unlabeled. For example, in the phrase *"Cash and cash equivalents,"* the ground truth treats as one entity, but the model only tagged "Cash", missing "cash equivalents". Subword tokenization further contributed to fragmentation, as seen with terms like "nonoperating" split into "non" and "operating," causing partial tagging (Joshi et al., 2020).

Span Boundary Errors and Overlapping Entities: The model's predicted entity spans often did not align precisely with true entity boundaries, including extra tokens or merging separate entities. In *"the notes bear interest at a stated rate of 6.25% per annum,"* the interest rate "6.25%" was incorrectly captured with "per annum", introducing false positives (Chen et al., 2022). Conversely, the model occasionally merged entities, such as *"net income in 2024,"* incorrectly combining distinct entities, reducing precision and recall (Yang et al., 2021).

Entity Type Confusions (Mislabeling): With 139 entity categories, the model frequently misassigned labels among semantically similar classes. For instance, a monetary quantity intended as Dividend might be mislabeled as Revenue, directly affecting precision and recall (Flowers, n.d.). These type-confusion errors, exacerbated by the dataset's fine-grained taxonomy, align with known challenges documented in financial NER research (Chen et al., 2022; Yang et al., 2021).

Comparison to Prior Work

Observed error patterns mirror those reported in related financial NER studies. Domain-specific models emphasize complex

terminology handling to reduce such errors (Chen et al., 2022; Yang et al., 2021). Our findings reinforce these known difficulties, highlighting challenges in fully capturing compound entities, precise boundary prediction, and fine-grained type differentiation (Joshi et al., 2020).

Observed diminishing returns in later epochs aligns with existing literature on transformer-based models, highlighting the potential need for innovative approaches such as curriculum learning, advanced data augmentation, or external knowledge integration. The final model's validation loss is of particular note, as the first epoch loss started considerably lower than previous model.

Improvements

Learning Rate

A reduced learning rate (e.g., 1e-5) prevents the model from making overly aggressive parameter updates, particularly in the early stages of training. This stability allows the optimizer to converge more gradually to a flatter minimum in the loss landscape—an outcome often associated with better generalization to unseen data (Devlin et al., 2019). In the context of financial NER, where certain entities are rare and easily overshadowed by more frequent classes, the slower update pace mitigates the risk of overfitting on dominant labels while enabling finer distinctions between semantically similar terms (Chen et al., 2022).

Warm-up Steps

Warm-up steps gradually increase the learning rate during the initial training phase, preventing the model from making destabilizing updates before the optimizer has a good sense of the loss surface. This has been shown to significantly reduce variance in gradient updates and improves convergence stability, especially in transformer-based models (Devlin et al., 2019). In our case, the warm-up mechanism contributed to lower initial validation loss and more effective training, as corroborated by the early epoch results.

Frequent Evaluation

Incorporating evaluation checkpoints during training allows for real-time tracking of model performance and early detection of overfitting or training stagnation. This was particularly helpful in our experiments where diminishing returns began after a few epochs. By monitoring metrics like F1-score, precision, and recall closely, we were able to understand when additional training stopped yielding improvements, which is in line with observations reported in prior transformer-based NER studies (Araci, 2019; Yang et al., 2021).

Conclusion

FinBERT's has significant potential in improving financial NER tasks. Substantial increases in precision, recall, and F1-score achieved through targeted preprocessing and hyperparameter tuning validate specialized transformer models' ability to classify financial sentiments. From the findings of earlier studies, methods to pursue future improvements include:

- **Span-based NER models**: Directly predicting entity spans might better capture multi-word entities and reduce fragmentation (Joshi et al., 2020).

- **Conditional Random Field (CRF) layer**: CRFs enforce valid label sequences, correcting inconsistent predictions and span overlaps (Lample et al., 2016).

- **Focal loss**: Applying focal loss can address class imbalance and emphasize learning from difficult, misclassified instances, potentially improving rare entity classification (Lin et al., 2017).

I hope to see more success in the field of Financial NER, with much room for development from entities with larger amounts of resources. There is potential for many possibilities, with finance consuming such a significant part of our lives.

References

- Araci, D. (2019). *FinBERT: Financial sentiment analysis with pre-trained language models*. arXiv preprint arXiv:1908.10063.
- Chen, X., Yang, J., Zheng, W., et al. (2022). *FinQANet: A pretrained language model for financial QA and NER*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).
- Flowers, J. G. (n.d.). *Financial-NER-NLP* [Data set]. Hugging Face. Retrieved February 26, 2025, from https://huggingface.co/datasets/Josephgflowers/Financial-NER-NLP
- Joshi, M., Levy, O., Zettlemoyer, L., & Weld, D. S. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8, 64–77. https://doi.org/10.1162/tacl_a_00300
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). *Focal loss for dense object detection*. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2980–2988. https://doi.org/10.1109/ICCV.2017.324
- Yang, Y., Zhang, Y., & Dong, D. (2021). *Finner: Finetuning BERT for financial named entity recognition*. Proceedings of the IEEE International Conference on Big Data (IEEE BigData).