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

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. The rapid advancement of NLP models, particularly transformer architectures such as BERT (Devlin et al., 2019), has provided promising avenues for automation. However, general-purpose NLP models often underperform in financial-specific contexts due to their inability to capture specialized jargon and nuanced language prevalent in financial texts. This study addresses these limitations by employing FinBERT (Araci, 2019), a financial-domain-specific model, exploring the accuracy and efficiency in financial entity extraction.

Background

Historically, financial NER was primarily addressed using rule-based methods and traditional machine learning techniques such as Conditional Random Fields, Support Vector Machines, and Hidden Markov Models. These models heavily depend on extensive feature engineering and domain-specific rules, limiting their scalability and generalizability.

Transformer-based architectures like BERT have significantly improved NLP tasks. 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.

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

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%.

Model Selection & Hyperparameter Tuning

Initially, a baseline model was fine-tuned using default FinBERT settings, achieving an F1-score of 48%. Several hyperparameter tuning strategies improved performance significantly. Lowering the learning rate to 1e-5 led to more stable training updates, reducing erratic fluctuations and consequently boosting precision from 42% to 56.17%. The batch size optimization at 32 ensured balanced updates, which mitigated issues of underfitting observed at smaller

batch sizes or instability with larger ones. 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

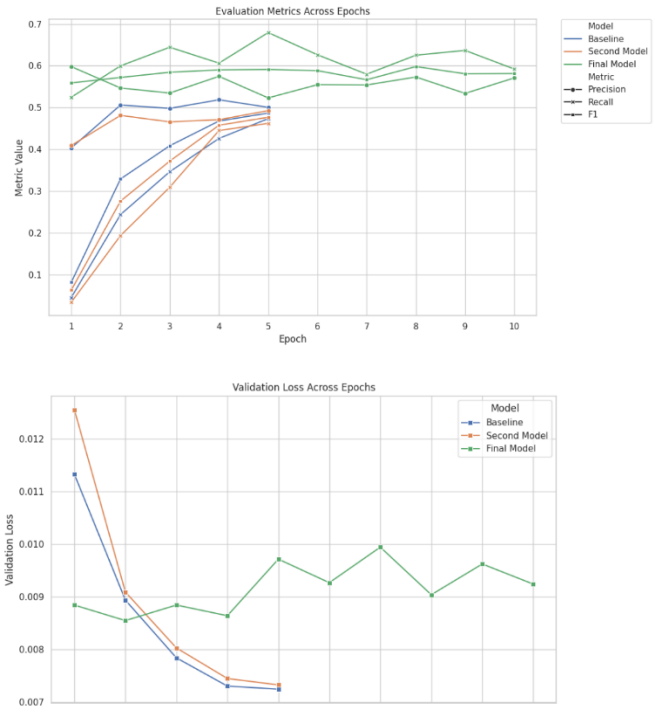

Fig. 2 Metric Improvement

Discussion

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. This could be attributed to the warm up steps, along with more accurate token alignment.

The iterated model's features that improved performance:

- Lower learning rate slowed down updates, allowing the model to better generalize to rare or noisy entities.

- Warm-up stabilized training in early steps, avoiding premature convergence.
- Frequent evaluation gave more insight into training dynamics and helped spot when performance plateaued or dipped.
- Increased epochs gave the model time to refine predictions beyond common tags.

Remaining challenges that persisted through the model improvements included:

- **Hyperparameter Sensitivity**: Lowering the learning rate improved training stability initially but may have led to insufficient model updates in later epochs, causing stagnation or slight regression in validation performance.
- **Complexity vs. Generalization**: The final model's increased complexity from extensive tuning might capture noise rather than generalizable patterns, leading to higher validation loss.
- **Early Stopping Consideration**: The earlier models might have had optimal performance due to implicit early stopping effects. Extending training in the final model without implementing explicit early stopping might have allowed minor overfitting.

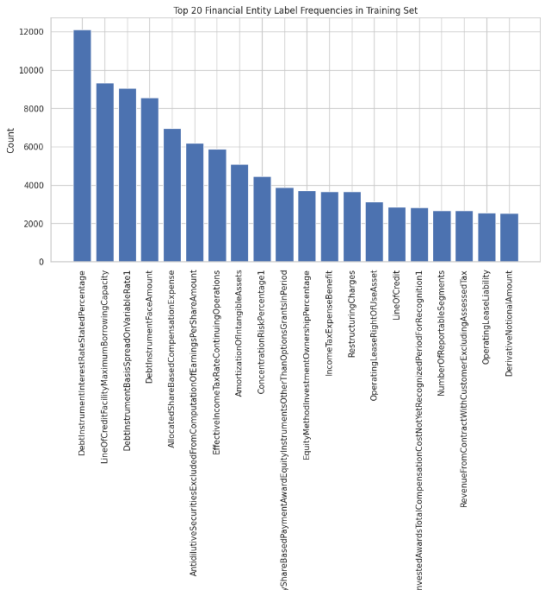

Fig 3. Class Distribution

The DebtInstrumentInterest-RateStatedPercentage label identifies the explicitly stated interest rate on a debt instrument such as a bond, loan, or note payable. It refers to the contractual rate specified in loan agreements or financial notes.

Example in text:

"The $500 million notes bear interest at a stated rate of 6.25% per annum."

Interest rates are a fundamental attribute of any debt instrument. Public companies are legally required to disclose this information in filings such as 10-Ks, quarterly reports, or bond issuance summaries. Because many disclosures include details about multiple debt instruments, this label appears frequently across documents. The LineOfCreditFacilityMaximumBorrowingCapacity label denotes the maximum amount a company can borrow under a revolving line of credit. This isn't necessarily the current debt outstanding, but the total

capacity available. Many companies maintain credit lines for liquidity purposes, and are required to disclose the capacity—even if they don't actively use it. These facilities are often renegotiated or renewed, leading to repeated mentions across time periods and reports. The DebtInstrument-BasisSpreadOnVariableRate1entity captures the spread (in basis points) added to a reference interest rate (like LIBOR or SOFR) to determine the variable interest rate for a debt instrument. Many corporate loans, especially revolving or term loans, use floating interest rates tied to benchmark rates. Disclosure of the spread is legally required and often appears alongside the base rate (e.g., "LIBOR + 2.5%"). The prevalence of these labels underscores the importance of interest-bearing instruments and credit arrangements in corporate financial disclosures. These labels not only appear frequently due to regulatory disclosure requirements but also represent critical inputs for financial modeling, valuation, and credit analysis.

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. Class imbalance and entity ambiguity remained significant challenges, emphasizing the importance of balanced training data and nuanced annotation methods.

From the findings of earlier studies, focal loss (Lin et al., 2017) downweights easy examples and focuses training on hard, misclassified tokens. This technique could be useful since label distribution is heavily skewed — many labels appear very rarely. This shifts attention to low-confidence predictions. Another method would be to implement Span-based NER. Token classification with IOB2 tagging can be fragile when labels are noisy or long entities are broken by subword tokenization (Joshi et al., 2020). This would reformulate the task as span classification where the model predicts (start, end, label) for entities. The goal would be to improve performance on longer, compound financial entities (e.g., ShareBasedCompensationArrangement...) that get fragmented in IOB tagging.

I hope to see more success in the field of Financial NER, with much room for development from entities with larger amounts of resources and longer time constraints.

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).