EFFICIENT FINANCIAL NAMED ENTITY RECOGNITION WITHOUT LARGE LANGUAGE MODELS

Hong Hu

hong.hu@berkeley.edu

ABSTRACT

This study explores the effectiveness of non-Large Language Models (LLMs) for Named Entity Recognition (NER) tasks in the financial industry, addressing concerns about data privacy and vendor updates associated with LLMs like GPT-4. Utilizing the FiNER-Open Research Dataset (FiNER-ORD), we implement baseline and improved models, comparing traditional and BERT-based approaches. Our findings reveal that a FastText-embedded LSTM achieves a baseline Entity F1 score of 0.72, which is surpassed by a fine-tuned BERT model yielding an F1 score of 0.86 through transfer learning from the CoNLL-2003 dataset. This model not only surpasses the LLM benchmarks but demonstrates the efficacy of tailored models in specific domains like financial NER, underscoring the potential of transfer learning in enhancing model performance without LLM reliance. The study also highlights the importance of careful tokenization strategies and shows that LSTM networks remain competitive for shorter sequence NER tasks.

1. INTRODUCTION

Information Extraction (IE) is a crucial component in the domain of Natural Language Processing (NLP), tasked with obtaining structured information from unstructured text data. Within IE, Named Entity Recognition (NER) plays a pivotal role by focusing on the identification and classification of entities, such as individuals, organizations, and locations, within a text, and subsequently assigning appropriate categorical types to them.

Though general-purpose LLMs, such as GPT-4, demonstrate satisfactory performance in Named Entity Recognition (NER), they present significant challenges in the financial industry due to concerns related to data privacy and the potential for unpredictable updates from vendors.

In this study, we investigate the application of NLP techniques to effectively conduct NER tasks without the reliance on LLMs, thus providing an alternative solution for financial institutions in need. Concurrently, we aim to explore methodologies that have the potential to exceed the performance of popular LLMs on existing financial NER benchmarks.

2. RELATED WORK

2.1. Large Language Models

Large Language Models (LLMs) have revolutionized the field of natural language processing by enabling significant advancements in understanding and generating human-like text. Recent models, such as GPT-4, utilize transformer architectures that scale up billions of parameters, allowing for improved context comprehension and language generation capabilities. These models are pretrained on massive corpora and fine-tuned for specific tasks, exhibiting state-of-the-art performance across various NLP benchmarks. Despite their remarkable achievements, LLMs pose challenges related to computational cost and ethical concerns, including biases and interpretability, which remain active areas of research and discussion within the community.

2.2. Financial Evaluation Benchmark

The Financial Evaluation Benchmark, as introduced by Shah et al. (2022), presents the inaugural heterogeneous evaluation framework known as FLUE, which encompasses five distinct financial NLP tasks. These tasks include financial sentiment analysis, news headline classification, named entity recognition (as explored by Alvarado et al., 2015), structural boundary detection, and question answering. Expanding upon this, FinBen emerges as a comprehensive open-source evaluation benchmark, incorporating a total of 36 datasets that cover 24 different financial tasks. Notably, it includes the two most prominent NER benchmarks within the financial sector. The performance metrics among various LLMs exhibit significant variation, with the state-of-the-art (SOTA) outcomes currently achieved by GPT-4, as illustrated in Table 1.

Dataset Metrics	Chat GPT	GPT 4	Gemini	FinMA 7B	Mixtral 7B
NER SEC Filling	0.77	**0.83**	0.61	0.69	0.24
FINER-ORD EntityF1	0.28	**0.77**	0.14	0	0.05

Table.1. Word-piece Tokenization.

3. DATA

The dataset employed for this study is the FiNER-Open Research Dataset (FiNER-ORD), which comprises 47,851 English financial news articles sourced from webz.io, as outlined by Shah et al. (2023). To facilitate the annotation of named entities, a subset of 220 articles was randomly selected, with 201 articles remaining post-filtering for vacant entries. The annotations were conducted using Doccano, an open-source annotation tool, where person (PER), location (LOC), and organization (ORG) entities were manually labeled. This annotation process involved multiple independent annotators to mitigate bias. FiNER-ORD was made available by the Fin AI working group, established with the objective of advancing open science, tooling, and model initiatives to ensure responsible innovation and application within financial services. The dataset is consolidated in a parquet file, which I manually divided into training, validation, and test sets with a distribution ratio of [70%, 15%, 15%]. FiNER-ORD represents a Named Entity Recognition (NER) dataset that was automatically compiled by implementing pattern-matching heuristics on financial news articles. An illustration of FiNER-ORD is presented below in Figure 1.

Jimmy Cao PER , a Beijing LOC -based BMW ORG spokesman, said he was not able to provide an immediate comment.

The average “maker suggested retail price” (MSRP) for all passenger cars remains relatively high in China LOC , at around 280,000 yuan ($45,000), according to research firm JATO Dynamics ORG .

Fig.1. Example of Annotation in FiNER-ORD.

The dataset utilizes the BIO tagging scheme, reminiscent of the CoNLL-2003 framework (Tjong Kim Sang & De Meulder, CoNLL 2003). Notably, the sole divergence from CoNLL-2003 is the absence of the miscellaneous (MISC) tag. It is important to note that the distribution of the three tags is not uniform. Consequently, to address this imbalance, oversampling techniques were employed prior to training. The implementation details involve grouping all sentences containing a single tag into three separate queues corresponding to PER, ORG, and LOC. If the LOC tags are undersampled, sentences with only one LOC tag are incrementally used to rectify the imbalance.

4. EXPERIEMENTS

In the initial phase of this study, the implementation of a baseline model was attempted using the Named Entity Recognition (NER) Tagging code provided in the Stanford CS230 Deep Learning course. This baseline approach involves constructing a vocabulary derived from the train, validation, and test datasets. Subsequently, each word is assigned a sequence identifier, denoted as the word index within the vocabulary. This sequence ID is transformed into an embedding by utilizing PyTorch's Embedding function. These embeddings are then processed through either a PyTorch Long Short-Term Memory (LSTM) network or a Transformer Encoder, culminating in a classification output layer that categorizes each word into one of the seven BIO tags.

However, it was soon observed that the vocabulary is confined to the current dataset. Consequently, the inclusion of additional datasets introduces new, unrecognized words into the vocabulary. Moreover, words that were unseen during the prediction phase are prone to misclassification. This observation highlights the necessity for developing an improved benchmark model that circumvents the limitations of a fixed vocabulary and efficiently accommodates unseen words during prediction.

Considering there will be multiple experiments, I run all experiments on a AWS g5.xlarge instance to reduce the training time. This instance is equipped with Nvidia A10G GPU with 24G additional Video Memory. The GPU Compute Capability is 8.6 as compared to 7.5 of commonly used Nvidia T4.

4.1. Benchmark

To simulate a real-world application, the decision was made to utilize FastText in conjunction with a PyTorch-based LSTM as the benchmark model. Unlike traditional word2vec models that fail to account for the morphological structure of words, FastText (Bojanowski et al., TACL 2017) represents each word as a collection of character n-grams. Each character n-gram is associated with a vector representation, and words are represented as the sum of these vectors. This approach facilitates rapid training on large corpora and enables the computation of word representations for out-of-vocabulary words. Furthermore, the FastText embedding method is efficient, achieving approximately 250 iterations per second during model training. This process is highly manageable, even on standard laptop computers.

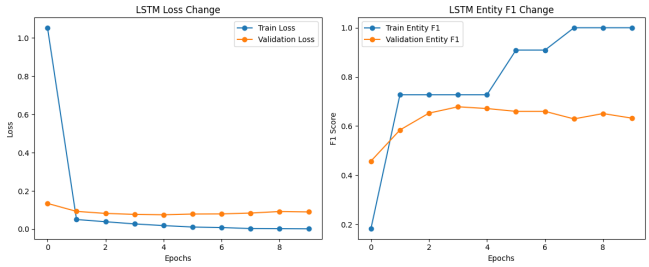

Fig.2. Training of Benchmark Model.

The loss history, illustrated in Figure 2, reflects the stability observed during the training process and highlights the constraints intrinsic to the benchmark model. One possible explanation for this is that the FastText embeddings are restricted to a dimensionality of 300, even when employing the comprehensive model, "cc.en.300.bin." Additionally, the PyTorch LSTM was initialized with random weights rather than with pre-trained parameters.

The evaluation of the model on the test dataset yielded an Entity F1 score of 0.72. As demonstrated in Table 2, this result underscores certain limitations inherent in the model's performance.

Model	Tag	TP	FP	FN	Sup	Precision	Recall	F1
LSTM	ALL	130	35	65	195	0.79	0.67	0.72
LSTM	ORG	52	16	50	102	0.76	0.51	0.61
LSTM	PER	37	5	7	44	0.88	0.84	0.86
LSTM	LOC	41	14	8	49	0.75	0.84	0.79

Table.2. Evaluation of Benchmark Model.

For instance, certain entities such as "Twitter" were not identified at all according to Figure 3.

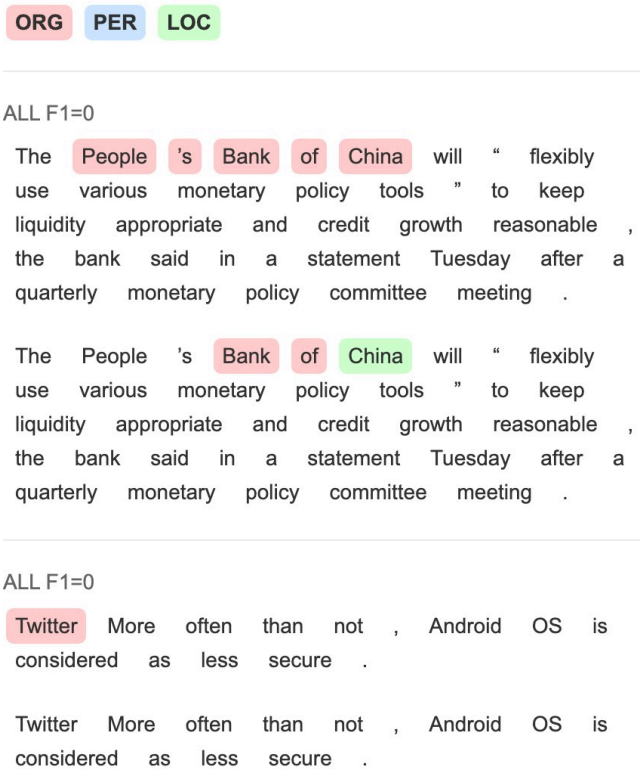

Fig.3. Misclassification of Benchmark Model.

4.2. Pretrained Transformers

In order to surpass the performance of the benchmark model, I have opted to enhance the embedding size and employ a fine-tuned token classification model. The use of BERT has become an apparent choice. The BERT model features 512 embeddings, enabling it to encapsulate more intricate semantic nuances. Besides, the BertForTokenClassification model is specifically fine-tuned for NER.

However, a challenge associated with employing BERT for NER tasks is its reliance on wordpiece tokenization rather than word tokenization. Consequently, it is necessary to define labels at the wordpiece level, as opposed to the word level. For instance, consider the two-word phrase "Eduardo DUHALDE," labeled as ["B-PER", "I-PER"]. Upon tokenization, it becomes four tokens, ['eduardo', 'du', '##hal', '##de']. Therefore, one must decide whether to propagate the original label of the word to all its wordpieces or to label only the initial wordpiece of each word while allowing the model to learn this pattern. As per the original BERT paper's methodology, I opted to label solely the first wordpiece, resulting in the labels ["B-PER", "I-PER", "NA", "NA"] for ['eduardo', 'du', '##hal', '##de']. The "NA" label is disregarded during the loss calculation. Additional information can be found in Table 3.

Token	Tag	Tag ID
eduardo	B-PER	4
du	I-PER	5
##hal	NA	-100
##de	NA	-100

Table.3. Word-piece Tokenization.

The performance evaluation, summarized in Table 4, indicates that the fine-tuned BERT model achieves parity with, or even surpasses, many LLMs.

Model	Tag	TP	FP	FN	Support	Precision	Recall	F1
BERT	ALL	167	45	28	195	0.79	0.86	0.82
BERT	ORG	81	31	21	102	0.72	0.79	0.76
BERT	PER	43	3	1	44	0.93	0.98	0.96
BERT	LOC	43	11	6	49	0.80	0.88	0.83

Table.4. Evaluation of Fine-tuned BERT Model.

Despite these advancements, certain challenges persist. For instance, when evaluating the entity "The Street Ratings" in Figure 4, the model is expected to accurately identify "The Street" as an Organization (ORG) entity. However, it often erroneously categorizes the complete phrase "The Street Ratings," which might relate to a magazine.

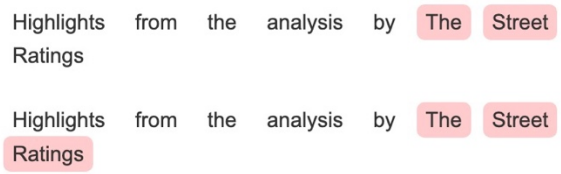

Fig.4. Misclassification of BERT Model.

The question arises as to whether there exists the potential for improvement in our approach. The answer is affirmative.

4.3. Transfer Learning

In our study, we have implemented the necessary modifications from the perspective of the model. However, further improvements can be achieved by enhancing the data side, specifically through the fine-tuning of the BERT model using additional Named Entity Recognition (NER) data and employing transfer learning techniques with the FiNER-ORD dataset. The CoNLL-2003 dataset is particularly suitable for transfer learning, as it comprises 14,041 labeled sentences, significantly surpassing the 752 training sentences available in FiNER-ORD. The process of transfer learning is relatively straightforward. Initially, the BERT model is trained using the CoNLL-2003 dataset and subsequently saved as a PyTorch tar file. This pretrained model is then loaded prior to training on the FiNER-ORD dataset. It is important to note that because our target dataset includes only ORG, PER, and LOC tags, it is necessary to convert the MISC tag to O.

The results indicate that transfer learning is effective. As demonstrated in Table 5, the model achieved an F1 score of 0.86, surpassing even that of GPT-4.

Model	Tag	TP	FP	FN	Support	Precision	Recall	F1
TRAN	ALL	173	33	22	195	0.84	0.89	0.86
TRAN	ORG	84	23	18	102	0.79	0.82	0.80
TRAN	PER	43	2	1	44	0.96	0.98	0.97
TRAN	LOC	46	8	3	49	0.85	0.94	0.89

Table.5. Evaluation of Transfer Learning Model.

Despite the continued challenge of accurately classifying ORG entities, where misclassifications can be made more effortlessly by humans, the efficacy of the model in Named Entity Recognition (NER) for personal (PER) entities is notably impressive, as illustrated in Figure 5. Only one PER entity was not identified, and one email address was mistakenly classified as a PER entity. The other misclassification remains ambiguous even to us. This showcases the efficacy of the transfer learning approach.

Fig.5. Misclassification of PER by Transfer-learning Model.

5. RESULTS AND DISCUSSION

Upon aggregating the results of the three models, it becomes evident that the BERT model initially enhances recall from 0.67 to 0.86. This improvement signifies that the fine-tuned BERT significantly aids in reducing classification omissions, characterized by fewer false negatives. Subsequently, transfer learning aids in elevating both precision and recall, notably enhancing precision by 5%. This indicates a reduced likelihood of misclassifying entities, evidenced by fewer false positives.

Model	Tag	TP	FP	FN	Support	Precision	Recall	F1
LSTM	ALL	130	35	65	195	0.79	0.67	0.72
BERT	ALL	167	45	28	195	0.79	0.86	0.82
TRAN	ALL	173	33	22	195	**0.84**	**0.89**	**0.86**

Table.4. Comparison of 3 Models.

A notable financial Named Entity Recognition (NER) dataset is derived from financial documents obtained from the U.S. Security and Exchange Commission (SEC) filings, commonly referred to as the FIN dataset (Salinas Alvarado et al., ALTA 2015). This collection encompasses sentences from public financial agreements submitted to the SEC, with entities manually annotated to include person (PER), location (LOC), and organization (ORG) types. The SEC filings demonstrate comparable performance when employing similar methodologies. However, a distinction in this study involved making minor modifications to fragment very lengthy sentences into multiple shorter ones, ensuring compliance with the maximum sequence length of 512, as dictated by PyTorch LSTM/Transformers, BERT, or other encoding frameworks. Furthermore, the SEC filings were pre-divided into training, validation, and test sets to facilitate experimental benchmarking. Given that the performance aligns with recent datasets yet originates from 2015 and is not as current as FiNER-ORD 2023, a detailed discussion was deemed unnecessary for the sake of conciseness.

6. CONCLUSIONS

This study demonstrates that, in the absence of Large Language Models (LLMs), a specialized model can still achieve competitive performance in the financial Named Entity Recognition (NER) domain by employing the appropriate embedding and encoder models, alongside strategies such as fine-tuning and transfer learning. It is essential, however, to exercise caution with WordPiece tokenization during BERT fine-tuning.

Another significant insight is that Long Short-Term Memory (LSTM) networks remain viable. Despite extensive hyperparameter tuning of the PyTorch TransformerEncoder, it fell short of surpassing the performance of PyTorch LSTM. This observation may be attributed to the nature of NER tasks, which typically involve sentences of considerably

shorter length compared to paragraph or document-level NLP tasks.

Additional incremental improvements were observed through techniques such as utilizing uncased embeddings, performing oversampling on training data, and employing the AdamW optimizer. While these approaches are common practices among NLP practitioners, their mention remains warranted due to their tangible contributions to model performance.

REFERENCES

FinBen: A Holistic Financial Benchmark for Large Language Models arXiv:2402.12659

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. WHEN FLUE MEETS FLANG: Benchmarks and Large Pre-trained Language Model for Financial Domain. arXiv preprint arXiv:2211.00083 (2022).

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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, Parramatta, Australia.