Information Extraction: An application to the domain of hyper-local financial data on developing countries

Anonymous Author(s)
Affiliation
Address
email

Abstract

Despite the need for financial and economic data in developing countries for development research and economic analysis, such data remains limited and siloed. In this paper, we develop and evaluate two Natural Language Processing (NLP) based approaches to address this issue. First, we curate a custom dataset specific to the domain of financial text data on developing countries and explore multiple approaches for information extraction. We then explore a text-to-text approach leveraging the transformer-based T5 model with the goal of undertaking simultaneous Named Entity Recognition and relation extraction. We find that this model is able to learn the custom text structure output data corresponding to the entities and their relations, resulting in an accuracy of 92.44%, a precision of 68.25% and a recall of 54.20% from our best T5 model on the combined task. Secondly, we explore an approach with sequential NER and relation extration. For the NER, we run pre-trained and fine-tuned models using SpaCy, and wee develop a custom relation extraction model using SpaCy's Dependency Parser output and some heuristics to determine entity relationships. We obtain an accuracy of 84.72%, a precision of 6.06% and a recall of 5.57% on this sequential task. Overall, LLMs such as T5 show tremendous promise in bridging the identified data gap.

1 Motivation and Related Work

Over 1 billion people around the world lack access to formal financial services and are either underbanked or unbanked [1]. In particular, the United Nations has identified financial inclusion as a direct enabler and target of at least 8 of the 17 Sustainable Development Goals[2][3]. Access to investment capital is also a significant issue in many parts of the world. Many activities such as progress monitoring of financial inclusion efforts, capital deployment to accelerate it, and economic impact assessment are highly dependent on the availability of hyper-local financial data in various countries. Such data is currently extremely limited in developing countries [4].

The use of natural language processing (NLP) techniques to extract and curate financial data on companies is a familiar one that is expected to accelerate due to advancements in large language models. Companies such as Bloomberg use information extraction techniques such as named entity recognition and relation extraction models to track organization activity such as how much capital companies have raised among other activities from a variety of online text sources [5]. Despite this progress, there are several limitations in the advancement of models for this type of application. For example, there is a gap in literature on annotated financial datasets for developing countries for Named Entity Recognition (NER) and relation extraction tasks. Second, there is a dearth of research on the performance of state-of-the-art NLP-based NER and information extraction models on financial data generated by non-Western countries. This is important because named entities and

Submitted to Computational Sustainability Workshop at NeurIPS 2023. Do not distribute.

financial news writing/reporting linguistic styles may vary significantly as we move from western English-speaking sources to non-Western/developing country sources.

1.1 Named Entity Recognition

Information extraction, task of identifying critical details like entity relationships from texts, has been extensively explored in NLP. This task can be unsupervised, semi-supervised, supervised or rule-based. Rule-based extraction, exemplified by works like Mykowiecka et al., relies on human-crafted rules[6]. Its limitations notwithstanding, it is widely applied in areas such as clinical data, with some studies like Hanafi et al. attempting automated rule creation[7]. On the other hand, supervised approaches, especially deep learning ones, have garnered attention for tasks such as extracting scientific information, as discussed in Han and Oleynik (2020)[8].

A central aspect of information extraction is Named Entity Recognition (NER). Exceptional performance in NER has been noted, with models like the one by Lample et al. achieving F-1 scores around 90% on English-NER[9]. Other works, like Chiu and Nichols (2016), also tout high performance with simpler architectures[10]. Notably, while supervised NER techniques dominate, there have been unsupervised learning approaches with varying success.

State-of-the-art performances in NLP are largely attributed to large language models such as GPT, T5, which are versatile across multiple NLP tasks[11]. While there are numerous examples of their applications in various domains, their utilization for simultaneous NER and relation extraction in finance has not been extensively documented.

Data availability is pivotal for model evaluation. Although datasets like CoNLL-2003 and OntoNotes-5 are staples for NER, finance-focused datasets, especially for developing countries, are scant[12][13].

Recognizing these gaps, our paper explores focuses on hyper-local financial data on developing countries and explores two extraction approaches:

- T5 LLM for simultaneous entity detection and relation extraction.
- A method based on a pre-trained CNN-based NER, Dependency Parsing, and heuristic extraction.

Our results, particularly with the T5 model, are promising and detailed in this paper's results section.

2 Approach

For our task, for a given sentence, we want the following:
Sample Input: "Apple had a net income of $9.4 million"
Sample Output: (Company:"Apple"), (Monetary_variable: "a net income"), (Monetary_value: "$9.4 million")

We explored two main approaches to this information extraction task: 1) a text-to-text t5 model for simultaneous entity extraction and relationship tags, and 2) a custom approach which combines outputs from a pre-trained NER model with those from a dependency parser with some heuristics.

2.1 Data

We scraped a prominent African financial news website (Tech Cabal) to retrieve 400+ news articles from 2019 and 2020 [14]. We then conducted other pre-processing steps on the dataset to change it to a format suitable for our models. We decided to structure our input at the paragraph-level instead of the sentence-level in order to capture relationships between entities that spanned several sentences. In order to use the T5 model, we manually labeled over 6,000 paragraphs in the following format:

company name, variable name, variable value, variable date

where variable name could be one of founder, country, revenue, customers/users, or investment.

In the cases where there were more than one piece of information in the text, we concatenated them all using a separator in the middle. The following shows one sample input and its corresponding target text:

Input text: f target text: *Jumia, revenue, €41 million, Q4 2020| Jumia, revenue, €33.7 million, Q3 2020|*

For our test set, we selected 20% of the training samples, ensuring no overlap with the training set to prevent model evaluation on memorized data. To assess model performance, we compared each word in the target output with its counterpart in the predicted output, employing both exact and fuzzy similarity scoring (above 90

2.2 LLM-based Simultaneous Entity Detection and Relation Extraction

For our t5-based model, we conducted three experiments:

- 1) Determining the best training set composition,
- 2) Evaluating the optimal model size, and
- 3) Hyper-parameter tuning.

Initially, we tested the small, base, and large t5 models for a single epoch. While there are bigger t5 models available, we could not use them due to our virtual machine constraints.

From the first experiment, optimal results came from training on the full set, then fine-tuning using a subset of informative samples balanced with non-informative ones.

The large t5 model stood out in performance, guiding our hyper-parameter adjustments. We varied batch size and epochs, keeping a consistent learning rate at 1e-4. Observing both the training and test set, we noted a stable 93% accuracy, even after 10 epochs spanning 5 hours. Pushing for 20 epochs with early stopping, however, resulted in overfitting and declined test performance.

2.3 Sequential CNN-based NER + Dependency Parsing and custom heuristics

Here, our aim was to tailor pre-trained NER and dependency parsing models to our specific task. We employed SpaCy - an open source model which uses residual convolutional neural networks and incremental parsing with bloom embeddings. Spacy's en_core_web_md" pipeline was used for primary entity detection and dependency parsing. With a focus on five entities: Companies, Persons, Country, Money, and Date, our goal was to align our outputs with the T5 model format.

Our extraction process concentrated on discerning the relationship between detected entities, primarily companies. For instance, a recognized Money entity would be categorized as 'investment' or 'revenue' if linked to a company. Similarly, a Person entity would be tagged as a 'founder' if related to the founding of the company.

We honed heuristics for these pairwise relations:

1. company - money
2. company - date
3. company - country
4. company - person.

To provide an example: in the statement, *"Apple had a net income of $9.4 million"*, the Money entity "$9.4 million" is connected to "Apple". To ascertain whether it is an investment or revenue, our classifier function uses embeddings of the phrase "a net income", comparing them to embeddings of revenue and investment-related words. With SpaCy's similarity function, we label the type if similarity surpasses 0.5.

We employed a similar approach for person entities and introduced a wrapper function for consolidating pairwise relations into a singular T5-compatible output.

Building on these foundational heuristics, we moved on to more intricate rules for relation detection using SpaCy's parser and NER tools. We iteratively refined our heuristics on a dataset subset and subsequently benchmarked against the full test data with T5.

Two distinct heuristics emerged: * Money-Company mapping:

- From a "MONEY" token, we probed its left ancestors for a related "ORGANIZATION".
- For absent subjects, we examined its verb's children for 'ORGS'.

* Company-Date mapping:

- For an "ORGANIZATION" token, we inspected neighboring prepositions for "DATE" entities.

Other detailed heuristics were also explored but not described here due to space constraints.

3 Results and Analysis

The following table summarizes the performance of our select models on the test set. We are only reporting the results from the evaluation based on exact matching.

Number	Approach	Model size	Epochs	Accuracy	Recall	Precision	Specificity	F1-score
1	Heuristic-Baseline	N/A	N/A	84.72%	5.57%	6.06%	92.03%	0.058
2	T5	Large	1	49.58%	55.99%	9.2%	48.98%	0.158
3	T5	Base	10	92.07%	67.41%	52.38%	94.34%	0.590
4	T5	Large	10	92.44%	68.25%	54.20%	94.68%	0.604
5	T5	Large	20	85.79%	32.09%	28.03%	91.40%	0.299

Our best model was the T5-based model run for 10 epochs.

Our T5-based model showcased commendable results on the test set, even with limited training data and exact matching criteria. We anticipate further improvements with more data. Upon comparing our T5 model's predictions with the baseline heuristic, we noticed that our evaluation may be overly rigorous. This stringency often penalized the model for minor deviations, like order or exactness.

For the baseline heuristic, it sometimes recognized company affiliates instead of founders, leading to undue penalties in our evaluation. The model's tendency to detect extra or fewer words also affected its score due to our strict evaluation parameters.

An assessment of both models unveiled issues in our data quality. There were instances where true relationships detected by the models were not labeled correctly in the training set. Enhancing our data quality would undoubtedly uplift both models' performance.

In conclusion, both methodologies present a promising avenue for information extraction in this context. Given the T5 model's effectiveness, we plan to integrate it into our main pipeline.

4 Conclusion and Future Work

In this paper, we have highlighted a gap in the use of NLP information extraction techniques on financial data originating from developing countries. We have curated a labelled dataset of entities and relations from over 6000 paragraphs of financial news sources from developing country sources. We have also presented two different approaches to this task of information extraction in this domain: one a staggered pre-trained CNN-based NER with dependency parsing and heuristics approach, and the other, a text-to-text approach to achieving this task using T5 - a transformer-based model by Google AI. We find that with the T5 model, despite only training on 6000 paragraphs, we achieved an F-1 score of 0.604 despite using stringent exact matching on a custom string output format. We obtained less impressive results with our best sequential NER-Dependency parsing + heuristic model when run on the full data, however, our analysis of why it scored poorly has yielded interesting directions which we will explore. We will also continue collecting more training examples in order to further fine-tune our T5 model-based approach, while also seeing if there are other transformer based approaches (e.g. BERT) that perform well on this task. Finally, beyond this paper, we aim to also explore a supervised relation extraction based approach in which a classifier model will be trained to classify the relationship between two named entities as accurate or inaccurate. It is our hope that efforts such as this can gradually help improve efforts to track financial inclusion in developing countries.

References

[1] Abraham Augustine. The limits of accelerating digital-only financial inclusion. Accessed: 2023-10-03.

[2] UNSGSA. Igniting sdg progress through digital financial inclusion. 2018. Accessed: 2023-10-03.

[3] John Kuada. Financial inclusion and the sustainable development goals. In *Extending financial inclusion in Africa*, pages 259–277. Elsevier, 2019.

[4] Alper Kara, Haoyong Zhou, and Yifan Zhou. Achieving the united nations' sustainable development goals through financial inclusion: A systematic literature review of access to finance across the globe. *International Review of Financial Analysis*, 77:101833, 2021.

[5] Fortune. How bloomberg used new language a.i. to 'tame the terminal', 6 2022. Accessed: 2023-10-03.

[6] Agnieszka Mykowiecka, Małgorzata Marciniak, and Anna Kupść. Rule-based information extraction from patients' clinical data. *Journal of biomedical informatics*, 42(5):923–936, 2009.

[7] Maeda F Hanafi, Azza Abouzied, Laura Chiticariu, and Yunyao Li. Seer: Auto-generating information extraction rules from user-specified examples. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6672–6682, 2017.

[8] Udo Hahn and Michel Oleynik. Medical information extraction in the age of deep learning. *Yearbook of Medical Informatics*, 29(1):208, 2020.

[9] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics.

[10] Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370, 2016.

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[12] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 2013.

[13] Erik F. Tjong Kim Sang and Fien De Meulder. 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, 2003.

[14] Tech cabal. `http://www.techcabal.com`. Accessed: 2020-02-20.