

EconBERTa: Towards Robust Extraction of Named Entities in Economics

Ashutosh Pathak
George Mason University
apathak2@gmu.edu

Chaithra Bekal
George Mason University
cbekal@gmu.edu

Vikas Velagapudi
George Mason University
vvelaga@gmu.edu

1 Introduction

1.1 Task / Research Question Description

The paper introduces EconBERTa, a language model pre-trained on economic texts, aimed at improving Named Entity Recognition (NER) in the economics literature, specifically for impact evaluation studies. The proposed model, EconBERTa, reaches state-of-the-art performance when trained upon the downstream NER task compared to other models such as BERT, DeBERTa, and RoBERTa. Additionally, the paper also analyzes the model’s generalization capacity, finding that the majority of the errors from the model are from a sub-span of features that the model was not able to understand. This paper also contributed a new dataset ECON-IE, which is a dataset of 1,000 abstracts from economic research annotated for entities describing the causal effects of policy interventions.

1.2 Motivation and Limitations of Existing Work

The field of computational linguistics has seen significant advancements with the adaptation of general-purpose language models for domain-specific tasks. The motivation for this model stems from similar previous works to adapt general-purpose language models using benchmarked datasets focused specifically on certain domains to improve the performance of downstream NER tasks in these domains. While models based on Science, Medicine, and Finance are built, this possibility has not been particularly explored in impact evaluation for Economics and this is the motivation to build the model.

The paper tries to overcome the limitations of the general language models by pretraining the model EconBERTa on a dataset specifically created for Economic papers and is termed Econ-IE.

They also focus on analyzing the weaknesses of a model. It remains unclear, which aspects of a model require improvement in order to increase robustness during deployment. This was also one of the motivations behind one of the tasks: analyzing the generalization capacity. The other papers did not analyze the errors to this extent and that was solved in this paper.

1.3 Proposed Approach

The proposed model in this paper involves pre-training and finetuning the models with ECON-IE dataset and comparing results. The dataset was collected and annotated for entities describing the causal effects of policy interventions: intervention, outcome, population, effect size, and coreference. After running the models on this annotated dataset, a thorough analysis was performed. The metric generally used in evaluating models is f1 score, but in this approach, the authors used Exact Match, Exact Boundary, Partial Match, and False Alarms which are entity relations.

In addition to just analyzing the results for accuracy, the authors also analyze if the model is just memorizing from the train set. They provide a clear distinction in Lexicosyntactic memorization and Lexical memorization.

1.4 Likely challenges and mitigations

The main challenge is ensuring the model’s robustness and its ability to generalize beyond memorized patterns. The paper is based on the fact of improving the performance of language models in the economic domain.

If reproduction is difficult, the plan includes examining error patterns and adjusting the model’s training process to improve its general-

ization capabilities. To this extent, the authors suggest fine-grained analyses and exploring different pretraining strategies to mitigate these errors from the model.

2 Related Work

There are several studies conducted to handle financial data using BERT-based architectures. Some of them referred by the authors were (Araci, 2019), (Yi Yang and Huang, 2020) and (Bo Peng and Huang, 2020). The paper (Araci, 2019) introduced FinBERT sentiment analysis, it was the first paper to introduce FinBERT.

In the paper (Yi Yang and Huang, 2020) the authors have performed fine-tuning over multiple Financial datasets using simple linear layer with a softmax activation function. They compare BERT and FinBERT performances over the different datasets to claim that FinBERT has better perform over Financial datasets. This was the second FinBERT model. As for (Bo Peng and Huang, 2020) their study was built on top of Yi Yang and Huang (2020). They compared the performances of FinBERT with the BERT on a wide variety of financial text processing tasks.

The first two papers stated here focused on Sentimental analysis while this one focused on evaluating our models on three semantic tasks that are more specific to the financial domain document causality detection (Dominique Mariko and de Mazancourt, 2020), numeral understanding (Chung-Chi Chen and Chen, 2019b), and numeral attachment (Chung-Chi Chen and Chen, 2020) along with the traditional sentiment analysis.

Our chosen paper works towards filling two gaps, It is the first paper to address information extraction from scientific economic content and also the first to define a NER annotation scheme and release an annotated dataset for causal entities of economic impact evaluation.

3 Experiments

3.1 Datasets

The ECON-IE Dataset was created for pre-training and fine-tuning EconBERTa and other models. It comprises 1000 abstracts from economics research papers, totaling over 7000

sentences. This dataset is publicly available via the authors' GitHub repository at https://github.com/worldbank/econberta-econie/tree/main/data/econ_ie

The data folder includes train, test, dev, and full .conll files, along with folders containing data used for cross-validation. The dataset was sampled from 10,000 studies curated by 3ie (Gaarder and White, 2009), published between 1990 and 2022, covering all 11 sectors defined by the World Bank Sector Taxonomy (Bumgarner, 2017). The authors employed stratified sampling to ensure diversity and generated a fixed held-out test set by sampling 20% of the abstracts; the remaining 80% were split for a 5-fold cross-validation set. We are using the train, dev, and test files for fine-tuning the models and comparing results.

3.2 Implementation

The central claim in this paper is that the developed model EconBERTa reaches state-of-the-art performance on their downstream NER task. The authors have compared EconBERTa's performances with BERT, RoBERTa and mDeBERTa-v3 models' performances on the given dataset. We defined code to fine-tune the models based on the fine-tuning parameters that were given in the paper to compare the performance as given in the paper. Our goal was to reproduce the results and prove/disprove their claim. We are using the train, dev, and test data files for our fine-tuning process. Initially we used the Word based approach for training where we used the dataset as given in the author's repository. After that we tried Sentence based approach since we did not see good results initially. This approach was better since the model was able to recognize context. Reproduction repository link: <https://github.com/pathak-ashutosh/EconBERTa-rep>

3.3 Results

In the Word based training approach we were not seeing good results and the reason was the model was not picking up the context of the sentence. We continued our work to figure this out and for the sentence based fine-tuning we got better results which were almost at par with the authors' results. We were able to prove their claim that EconBERTa is the state-of-the-art model for Economic texts. Given below is the table comparing the f1 scores

of the models we fine-tuned

Table 1: Model Performance Comparison

Model	Authors' F1-scores
EconBERTa-FC	0.687
mDeBERTa-V3	0.670
RoBERTa	0.659
BERT	0.649

Table 2: Model Performance Comparison

Model/F1-score	Word Based	Sentence-Based
EconBERTa-FC	0.16	0.6778
mDeBERTa-V3	0.95	0.6253
RoBERTa	0.42	0.5566
BERT	0.3832	0.4857

3.4 Discussion

We initially performed fine-tuning using just one out of three learning rates (5e-05) since it requires a lot of time for training. While performing reproduction of results on the given dataset, we faced many challenges, one of them was preprocessing the data. The annotation guidelines are complex and the labels were not as simple as giving positive, negative or neutral. There were a couple of hiccups while working on getting the dataset into batches to train the model.

The authors have not given enough information in the paper about the preprocessing. Since, this is a new dataset developed by the authors, additional information on preprocessing would make reproducing easier. In addition to that, their implementation uses AllenNLP which is totally different from our implementation in PyTorch which could also be one of the reasons we were not able to verify their claim in the first attempt.

But changing our approach to train the data as sentences helped us capture the context and give expected results.

3.5 Resources

Replicating models like BERT, RoBERTa, DeBERTa, and EconBERTa demands significant computational resources, accessed through GMU's GPUs. After changing our approach training each epoch took about 5 minutes, posing lesser time constraints, we were able to try out a

couple of learning rates as well.

We developed code from scratch using PyTorch to integrate with Huggingface Transformers, enabling us to train and test models against the provided dataset, aiming for results matching the original study. This phase required advanced programming skills and deep model understanding. Additionally, this process involved rigorous debugging, optimization, and experimentation to achieve competitive performance metrics comparable to the state-of-the-art results reported in the literature.

3.6 Error Analysis

In our efforts to replicate the results reported by the authors in their model comparisons, we encountered significant discrepancies earlier between the expected and observed performances of the models. But we were able to fix the approach as explained earlier and get better results. We performed error analysis and plotted the graph (Figure 1) to see the proportion of Exact Match, Partial Match and other metrics proposed in the paper. The authors perform error analysis with respect to Lexical and Lexicosyntactic Memorization.

A deeper examination into the annotations and the lexicological changes proposed by the author reveals potential sources of these errors. The introduction of entity-level annotations aimed to enhance model performance by more precisely identifying entity types. However, our analysis identified significant inconsistencies in how these annotations are applied and we can see those differences in Figure 1.

4 Checkpoint 2 Approaches

4.1 Robustness Approach

4.1.1 Approach for Evaluating the Model Robustness

In this checkpoint, we focused on evaluating the robustness of EconBERTa by introducing various forms of data perturbations. We created synthetic perturbation data by introducing noise such as typos. This perturbation data served as a robustness benchmark. We evaluated the model using metrics such as F1 score across the perturbed data to assess how well the model could

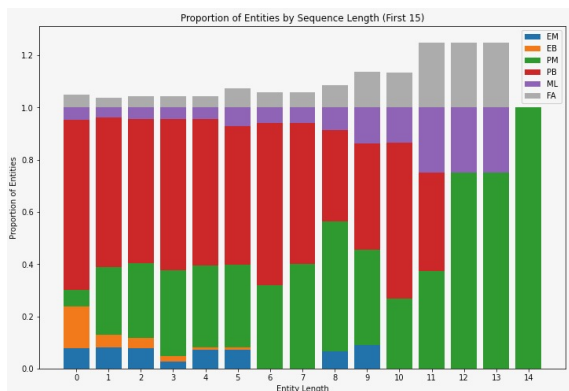


Figure 1: Proportion of exact matches and of the different error types for EconBERTa

maintain its performance under less-than-ideal conditions.

We also performed Tests given in the checklist paper. The tests defined in checklist paper were mainly MFT(Minimum Functionality Test), INV(Invariance) and DIR(Directional Expectation Test). For MFT tests our goal was to verify how the model behaves when we swap similar entities. Can it figure out that the words that replaced the existing entity still belong to the same entity? In the INV tests, we focused on changing the casing of the entities and then getting the model to predict the same entity type, essentially verifying case sensitivity. When it comes to DIR tests, we tested model’s robustness by introducing typos, changing numerical values, and swapping punctuations.

4.1.2 Approach for Improving the Model Robustness

To enhance the robustness of EconBERTa, we adjusted our training strategy by incorporating adversarial training techniques. We generated adversarial examples during the training phase to make the model less sensitive to small perturbations in the input data. We used a mix of original and perturbed data for training, with a validation set kept strictly from the original unperturbed data to monitor generalization. The model was trained using a learning rate of $7e-5$. We created some examples out of the unseen test data to test the model after retraining.

4.2 Multilinguality Approach

Our approach to developing a multilingual version of EconBERTa involved training the model with

a dataset comprising economic texts in Spanish language Spanish. We utilized machine-translated versions of the ECON-IE dataset to create training and validation datasets. The model was trained using the same hyperparameters as the monolingual model to maintain consistency in the training process.

5 Checkpoint 2 Experimental Results

5.1 Results of Robustness Evaluation

The evaluation of EconBERTa on the robustness benchmark showed mixed results. There was a slight decrease in the F1 score after introducing the perturbation. The model performed well in cases where perturbations were minor (e.g., slight spelling errors), but struggled with more significant disruptions like intense typos. For example, in cases of minor typos, the model maintained an F1 score above 0.65, but in cases of higher typo rate, the score dropped to around 0.53. But this seems expected since adding heavy typos will change the meaning of the words and overall sentence.

We performed MFT tests on EconBERTa model for which we observed good results [Figure2]. The Minimum Functionality Test results indicate a disparity in the performance of the evaluated system or algorithm across various recognition tasks. Notably, the system exhibits commendable accuracy in coreference recognition, with a minimal failure rate of 1.8%, suggesting a high level of proficiency in identifying and linking relevant entities within the text. Similarly, the system demonstrates strong capabilities in population recognition and intervention recognition, with failure rates of 3.7% and 2.0% respectively, indicating reliable performance in these areas. Conversely, the system’s performance in outcome recognition presents a significant challenge, as evidenced by a higher failure rate of 9.5%, which may necessitate targeted improvements. The effect-size recognition task also shows room for enhancement with a failure rate of 5.3%. These insights highlight specific areas where the system excels and where it could benefit from further refinement, providing a clear direction for future development efforts to bolster overall performance.

Upon performing INV tests [Figure3] related






MINIMUM FUNCTIONALITY TEST		
	test name	failure rate
+	Test for correct population recognition	7 / 187 = 3.7% 
+	Test for correct intervention recognition	5 / 250 = 2.0% 
+	Test for correct outcome recognition	34 / 357 = 9.5% 
+	Test for correct effect-size recognition	5 / 94 = 5.3% 
+	Test for correct coreference recognition	2 / 112 = 1.8% 

Figure 2: MFT Results






INVARIANCE TEST		
	test name	failure rate
	Test for intervention invariance	0 / 250 = 0.0% 
	Test for population invariance	0 / 187 = 0.0% 
	Test for outcome invariance	0 / 357 = 0.0% 
	Test for effect_size invariance	0 / 94 = 0.0% 
	Test for coreference invariance	0 / 112 = 0.0% 

Figure 3: INV Results

to case sensitivity, we found out that the failure rate was 0. This means that the model does not distinguish between upper-cased entities with lower-cased ones. It can clearly identify and tag these entities with 100% accuracy.

5.2 Results of Improved Model

The improved model showed similar results on the perturbation benchmark, with an average F1 score almost same after perturbations compared to the original model. However, the performance on the original test set remained consistent, indicating that the model’s ability to generalize was not compromised by the robustness enhancements. Due to lesser availability of time, we retrained the model only with one type of perturbation. As a future work we can also train with more such perturbed data and observe better robustness.

5.3 Multilingual Model Performance

In this checkpoint, we delved into the challenges of applying NLP models to multilingual datasets by translating an English dataset into Spanish. Our results highlighted a notable decline in performance across all models, which underscores the difficulties of cross-linguistic model application. EconBERTa led the pack with the highest F1 score of 0.2966, followed closely by mDeBERTa, RoBERTa, and BERT multilingual. Table 3 gives the different F1-scores for multilingual models. This significant drop in scores likely stems from changes in sentence structure and the challenges of maintaining semantic integrity during translation. Our annotations were specifically made

by Experts who did them for the english dataset depriving us of a valuable step when trying to replicate this in the Spanish set.

Regarding the English dataset, our analysis did show that EconBERTa outperformed the other models, consistent with the original claims of its superior effectiveness. Although we couldn’t replicate the exact F1 scores from the study, the relative performance trends were similar. This suggests that while EconBERTa may have a competitive edge in English, applying these models directly to another language without addressing linguistic differences results in less than ideal performance. This experiment underscores the importance of either developing robust multilingual models or tailoring training methods to better accommodate different linguistic contexts, enhancing the effectiveness of NLP models across various languages.

Table 3: Multilingual Performance

Model/F1 Score	English Dataset	Spanish Dataset
EconBERTa-FC	0.6778	0.2966
mDeBERTa-V3	0.6253	0.2804
XLNet-RoBERTa	0.5366	0.2770
BERT multilingual	0.4857	0.2631

6 Discussions

Throughout the project, we encountered significant challenges in replicating the original model’s results, primarily due to the complex preprocessing required by the ECON-IE dataset and the lack of detailed preprocessing guidance in the original papers. This led to discrepancies in model performance during our initial replication attempts. Additionally, the replication process was complicated by differences in the implementation frameworks (AllenNLP vs. PyTorch) and potential undisclosed pre-training steps by the original authors, which contributed to significant discrepancies observed in the F1 scores. These challenges underscore the importance of transparency and detailed documentation in machine learning projects to ensure reproducibility and consistency across different replication attempts.

Moreover, our efforts to enhance the model’s robustness through data perturbations and adversarial training techniques revealed that while the model could handle minor perturbations, its

performance was significantly impacted by more substantial disruptions, highlighting the need for continuous improvements in model robustness. The extension to multilingual capabilities further emphasized the difficulties of applying NLP models across different languages, as evidenced by the performance drop when applying the model to the Spanish dataset. This suggests a pressing need for developing robust multilingual models or tailoring training methods to better accommodate linguistic differences, ensuring effective model performance across various languages.

7 Workload Clarification

The workload for both checkpoints was evenly distributed among team members. As a team, we read the paper first and then we decided to work on the code together as we wanted to implement fine-tuning for the reproduction. We divided the functions in the python code shared in our repository and fixed errors as and when we got them. We finally ran the models by dividing BERT, RoBERTa, mDeBERTa-v3 and EconBERTa among ourselves and obtained results to save computational time. In the report we divided the sections and completed them accordingly. Each member focused on different aspects of the checkpoint 2 including testing robustness, improving the model and multilinguality to ensure an equitable distribution of tasks and efficient project progression.

8 Conclusion

In conclusion, our efforts in checkpoint 1 demonstrated challenges in reproducing the original model's results initially, primarily due to discrepancies in data loading and preprocessing. In checkpoint 2, we first managed to address the discrepancies from checkpoint 1 and then successfully enhanced the model's robustness and extended its capabilities to understand multiple languages. While the multilingual model shows promise, further refinements are needed to achieve optimal performance across all languages. Future work could focus on improving data quality for training and exploring more advanced multilingual training techniques.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. In *arXiv preprint arXiv: 1908.10063*.

Yu-Yin Hsu Bo Peng, Emmanuele Chersoni and Churen Huang. 2020. Is domain adaptation worth your investment? comparing bert and finbert on financial tasks. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, page 37–44.

Kimberly Marie Bumgarner. 2017. [New sector taxonomy and definitions](#).

Hiroya Takamura Chung-Chi Chen, Hen-Hsen Huang and Hsin-Hsi Chen. 2019b. Overview of the ntcir14 finnum task: Fine-grained numeral understanding in financial social media data. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, page 19–27.

Hiroya Takamura Chung-Chi Chen, Hen-Hsen Huang and Hsin-Hsi Chen. 2020. Overview of the ntcir15 finnum-2 task: Numeral attachment in financial tweets. In *Development*, 850(194):1–044.

Yagmur Ozturk Hanna Abi Akl Dominique Mariko, Estelle Labidurie and Hugues de Mazancourt. 2020. Data processing and annotation schemes for fincausal shared task. In *arXiv preprint arXiv:2012.02498*.

Marie Gaarder and Howard White. 2009. [The international initiative for impact evaluation \(3ie\): an introduction](#). *Journal of Development Effectiveness*, 1(3):378–386.

Mark Christopher Siy Uy Yi Yang and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. In *arXiv preprint arXiv:2006.08097*.