CLRG@FinCausal2025: Cause-Effect Extraction in Finance Domain

Vibhavkrishnan K S, Pattabhi RK Rao and Sobha Lalitha Devi
AU-KBC Research Centre,
MIT Campus of Anna University, Chennai, India
sobha@au-kbc.org

Abstract

Thispaper presents our work on Cause-Effect information extraction specifically in the financial domain. Cause and effect information is very much needed for expert decision making. Particularly, in the financial domain, the fund managers, financial analysts, etc. need to have the he information on cause-effects for their works. Natural Language Processing (NLP)techniques help in the automatic extraction of cause and effect from a given text. In this work, we buildvarious cause-effect text span detection models using pre-trained transformer-based language models and fine tune these models using the data provided by FinCausal 2025 task organizers. We have only usedFinCausal 2025data sets to train our models. No other external data is used. Our ensemble of sequence tagging models based on theFine-tunedRoBERTa-Large language model achieves SAS score of 0.9604 and Exact match score of 0.7214 for English. Similarly for Spanish we obtain SAS score of 0.9607 and Exact match score of 0.7166. This is our first time participation in the FinCausal 2025 Task.

1 Introduction

Domain-specific causal information is very important for an informed decision making, particularly in expert decision-making processes. For example, financial organizations collect historical data of stock price movements and their causes to develop effective trading strategies.

Financial institutes collect and store causality information in English and other languages to understand early stock price fluctuation. The required information is published in different forms in different languages and magazines. All these information needs to be processed in real time for it to be useful for any decision making. Therefore, there is a need to develop automatic cause-effect information extraction systems.

The FinCausal2025 shared task at the Financial Narrative Processing Workshop (FNP) addresses this step by providing annotated data in English and Spanish. This paper further describes our work on the participation in this FinCausal 2025 shared task where we have developed span based models by fine tuning pre-trained large language models for our purpose.

2 Related work

The goal of the Fin Causal 2025 shared work (Moreno et al., 2025) was to identify causation in financial records. It was headed by Antonio Moreno Sandoval, Blanca Carbajo Coronado, JordiPortaZamorano, Yanco Amor TorteroloOrta, and DoaaSamy. This version analyzed datasets selected from English and Spanish annual reports, signaling a move away from extractive approaches and toward question-answering (QA)-focused strategies. Semantic Answer Similarity (SAS) and Exact Match (EM), two assessment measures, were highlighted in the challenge, along with abstractive question design. Advanced transformer-based models were utilized by the participants, and performance was improved by strategies such multilingual datasets and LoRA fine-tuning.

Dominique Mariko, Mahmoud El-Haj, and his team lead the Fin Causal 2023 shared task, which provided improved English and Spanish datasets with complex causal structures, including multi-effect causes and multi-cause effects. Robust system assessment was achieved by using evaluation criteria such as token-level F1 scores and Exact Match. Innovative techniques including retrieval-augmented generation and chain-of-thought prompting, together with state-of-the-art models like RoBERTa, Span BERT, and GPT-4-based architectures, were used by teams to push

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 236–241

the limits of causality identification in multilingual environments.

Building on previous iterations, the Fin Causal 2022 joint effort, headed by Dominique Mariko, Kim Trottier, and Mahmoud El-Haj, concentrated solely on causality detection. Financial news from 2019 and excerpts from SEC filings were added to the dataset. With the goal of identifying causes and effects in financial texts, participants made significant progress in detecting causality. Team SPOCK outperformed the other contestants in the use of ensemble sequence tagging models with RoBERTa-Large and the BIO scheme. Other noteworthy contributions were iLab's graph-based embeddings and Expert Neurons' clever pre-processing algorithms, which demonstrated a variety of approaches to successfully address causality extraction.

By supplementing the dataset with more instances from financial news, the Fin Causal 2021 shared task—which was managed by Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, StephaneDurfort, Hugues de Mazancourt, and Mahmoud El-Haj—further improved causality extraction. NUS-IDS, the victorious team, used a BERT-CRF in conjunction with a Viterbi decoder, using dependency graphs for token categorization. To get high accuracy in identifying causal sequences, other groups tried ensemble learning, sequence labeling, and graph neural networks. Even with improvements, there were still several difficulties, such as forecasting intricate causal networks, which highlights the need for more research.

The topic of causality identification in financial narratives has grown as a result of these common objectives, showing how methods have developed from straightforward extraction to complex, context-aware generative models and multi-layered analytical frameworks.

3 System Description

Our model makes use of the XLM-RoBERTa architecture, which is ideal for multilingual question-answering tasks since it uses self-attention methods to record contextual dependencies. The fundamental concept behind improving the model is to apply it directly to the span-based answer prediction problem, which entails guessing the beginning and ending locations of a response in the context. In order to comprehend and interpret the context effectively, this transformer network-based model framework functions inside a strong self-attention mechanism (Conneau et al., 2020).

$$L(\theta) = -\sum \log P(y_i|x_i, \theta) \quad (1)$$

where y_i represents the correct answer span, x_i is the context, and P(yi|xi,θ) is the predicted probability for the answer span (Devlin et al., 2018).

In addition to this, the model's training involves minimizing the **span loss**, which is designed to optimize both the start and end positions of the answer span. The span loss can be represented as:

$$L_{span}(\theta) = \alpha . L_{start}(\theta) + \beta . L_{end}(\theta) \quad (2)$$

where $L_{start}(\theta)$ is the loss for the predicted start position, $L_{end}(\theta)$ is the loss for the predicted end position, and α (alpha) and β (beta) are weighting factors to balance the start and end position contributions.

The model's performance is evaluated using two main metrics: the **Span Answering Score (SAS)** and **Exact Match (EM)**. SAS evaluates the semantic correctness of the predicted answer span in relation to the true answer, considering not just the overlap but also the meaning captured in the span.These metrics provide a comprehensive evaluation of both the relevance (SAS) and precision (EM) of the model's predictions.

3.1 Models

We used four models in our study, all of which were built on the XLM-RoBERTa architecture, which works well for multilingual question-answering tasks. Adapted to the Squad format for span-based answer prediction, these models comprise the conventional pre-trained XLM-RoBERTa base model (Conneau et al., 2020) and refined versions of the XLM-RoBERTa base and big models. We employed the following models:

a) **Standard XLM-RoBERTa Base (Squad):** This is the pre-trained, standard XLM-RoBERTa base model that has been optimized for question-answering tasks using the Squad dataset.

b) **Fine-Tuned XLM-RoBERTa Base (Squad):** This version improves on the previously trained base model by adding

optimized hyperparameters and fine-tuning it using our unique training data.

c) **Normal XLM-RoBERTa Large (Squad):** This large form of XLM-RoBERTa is pre-trained on Squad and provides a greater capacity for learning from data.
d) **Fine-Tuned XLM-RoBERTa Large (Squad):** This model combines changes to the learning rate, batch size, and epochs, and is based on the large version of XLM-RoBERTa that has been adjusted using our data.

We did not change the model architecture or add any additional parameters for fine-tuning. To enhance performance for the question-answering task, we instead changed the training parameters, including the learning rate, batch size, and number of epochs. The model's pre-existing parameters were refined throughout this fine-tuning procedure, which improved the model's fit to our particular dataset. Using the training code, which analyzes the input data (questions and situations) and modifies the start and finish locations of responses according to the tokenized outputs, the models were improved.

The table below contains the parameters for each model that was utilized. These provide information on the training parameters, model size, and particular fine-tuning techniques used.

Model Name	Pre-Trained Parameters
XLM-Roberta-Base-Squad2	279M
XLM-Roberta-Large-Squad2	550M

Table 1. Parameters of Models used

4 Training Process

4.1 Dataset

The financial text data in the dataset we got was organized in a CSV format and included the following columns: ID, Text, Question, and Answer. We updated the Answer column to incorporate the specific data required for span-based predictions in order to modify the data for optimizing our question-answering model. To be more precise, we transformed the response field into a JSON-like format that included the response text and the context's start and end indices. This made it possible for the model to pinpoint the precise place of the response within the given context.

For example, consider the following modification from the dataset **Original**:

- **Context**: "Nationwide is in robust financial health, having achieved profits of over Â£1 billion for the third consecutive year. As a mutual, profits are not the only barometer of our success, but they are important becausethey allow us to maintain our financial strength, to invest with confidence, and to return value to you, our members, through pricing and service."
- **Question**: "What is the effect of achieving profits of over £1 billion for the third consecutive year?"
- **Answers**: {"text": ["Nationwide is in robust financial health"], "answer_start": [0], "answer end": [40]}

Effective training and precise question-answering on financial data were made possible by the transformation we carried out, which guaranteed the model could read the precise answer span inside the surrounding text.

4.2 Hyperparameter Fine Tuning

In our approach for fine-tuning XLM-RoBERTawe follow on the work of (Moraites et al., 2021, Wolf et al, 2019), who offered a thorough framework for training subject classification models with Hugging Face's Transformers library. Although their configuration provided a strong basis for training the model, we modified it to better fit the particulars of our financial dataset. Increasing the number of epochs from the initial setting to seven was a crucial change that enabled the model to go through more thorough training and better absorb the subtleties of the financial data. In order to achieve effective gradient descent during training and maximize the trade-off between stability and quick convergence, we also changed the learning rate to 5e-5. Refining the batch sizes was another important modification. We set the evaluation batch size at 64 and the per-device training batch size at 16. These modifications were designed to ensure adequate data flow for model learning while managing memory limitations on our hardware. In order to avoid over fitting, we also adjusted regularization parameters like the weight decay (set at 0.01) and

added warmup steps (500) to progressively raise the learning rate during the first training phases. The model's efficiency and generalization to the financial question-answering tasks were enhanced by these adjusted parameters in conjunction with meticulous monitoring of training and evaluation performance.

5 Results and Discussion

Table 2 and 3presents a summary of our trials, comparing the performance of XLM-RoBERTa Base and Large models across Practice and Development datasets with and without fine-tuning. Exact Match (EM), which assesses exact token-level matches, and Semantic Answer Similarity (SAS), which measures semantic alignment between predictions and ground truth, are important assessment metrics. These tests are conducted for both Spanish and English datasets, demonstrating the models' multilingualism.

Using their respective Development datasets, the English and Spanish datasets underwent independent fine-tuning procedures. By taking use of the unique traits and subtleties of the English and Spanish environments, this guarantees that the models were tuned separately for each language.

The outcomes repeatedly show that model performance is much improved by fine-tuning. In every measure and language, fine-tuned models perform better than their non-fine-tuned counterparts for the Practice and Development datasets. Interestingly, EM scores demonstrate significant increases, especially in Spanish datasets, with gains of more than 50 percentage points in certain cases, while SAS scores for fine-tuned models routinely above 0.90 in the majority of setups.

Fine-tuned XLM-RoBERTa-Large demonstrates its outstanding ability to comprehend semantics by achieving the highest SAS score of 0.96 on the Practice dataset in English datasets. The Large model consistently demonstrates its capacity to generalize between phases on the Development dataset, attaining an EM score of 0.61 and an SAS score of 0.91. The Base model receives comparable scores, with an EM of 0.70 and an SAS of 0.94 on the Development dataset, although trailing the large model by a little margin in SAS. While the Base model offers a compromise between semantic comprehension and accuracy in some contexts, our results highlight the large model's superiority in managing semantic complexity.

Spanish datasets show that fine-tuning has a major effect, especially on Exact Match scores. After fine-tuning, for example, the EM of the Base model on the Practice dataset increases from 0.13 to 0.73. With the EM score increasing from 0.17 to 0.71 on the Development dataset, the refined Base model displays a comparable pattern. The fine-tuned large model achieved a peak SAS of 0.96 on the Practice dataset, and similarly, the fine-tuned models' SAS scores above 0.95 on both datasets. These findings show that the models can successfully adjust to multilingual data, particularly in Spanish and highlight the significance of fine-tuning in improving performance across both SAS and EM measures.

These findings provide several insights:

a) Making adjustments to language-specific the significance of adapting the models to the language and contextual peculiarities of English and Spanish is shown in the necessity of development datasets for optimizing SAS and EM scores.
b) The Base model's success in EM demonstrates its computational economy, while the XLM-RoBERTa-Large model's superiority in SAS qualifies it for semantically rich jobs.

Spanish datasets highlight the difficulty of multilingual adaptation by relying more on fine-tuning for better performance.

5.1 Performance of Testing Dataset

Following fine-tuning, both the English and Spanish dataset's performance on the Testing dataset exhibits notable gains. Semantic Answer Similarity (SAS) for English shows significant improvements with refined models, as the Base model rises from 0.73 to 0.93 and the large model rises from 0.78 to 0.96. Exact Match (EM) scores also increase, rising from 0.21 to 0.68 for the Base model and from 0.28 to 0.72 for the large model. Likewise, with the Spanish dataset, the large model achieves 0.96 for SAS and 0.71 for EM, while the Base model's SAS and EM improve from 0.76 to 0.96 and 0.16 to 0.76, respectively. These outcomes highlight the effectiveness of fine-tuning. Results from the Testing dataset will be incorporated into future research to provide a more thorough assessment of the models' generalization ability. The Testing dataset provides an objective assessment of the

models' performance on unknown data, whereas the Practice and Development datasets concentrate on training and fine-tuning. This stage is crucial for evaluating their robustness and real-world application, making sure they can correctly forecast responses in a variety of situations. These assessments will round out the conversation and provide more in-depth understanding of the model's performance.

5.2 Comparison to other systems

Comparing our study to other participating systems, we obtained competitive findings. Our algorithm performed well on a variety of datasets and came in at number four overall. Interestingly, our method performed well on some datasets, even though the best-performing system often produced better results. This demonstrates how well our system works in specific situations and emphasizes how flexible it is with regard to various kinds of data. A more thorough analysis of the variables influencing these variations, such as model setups, dataset management, and fine-tuning strategies, may yield insightful information for future system improvement and comprehension of its advantages and disadvantages.

Acknowledgments

We thank FinCausal 2025 organizers for providing the datasets and giving all the support in participating in the FinCausal 2025 task

References

Alexis Conneau, KartikayKhandelwal, NamanGoyal, VishravChaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, MyleOtt, Luke Zettlemoyer, and VeselinStoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440– 8451.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). HuggingFace's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

A. Moreno-Sandoval, J. Porta-Zamorano, B. Carbajo-Coronado, D. Samy, D. Mariko, and M. El-Haj. 2023. The Financial Document Causality Detection Shared Task (FinCausal 2023). In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), pages 2855–2860. IEEE, Sorrento, Italy. DOI: 10.1109/BigData59044.2023.10386745.

D. Mariko, H. Abi-Akl, K. Trottier, and M. El-Haj. 2022. The Financial Causality Extraction Shared Task (FinCausal 2022). In Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022, pages 105–107. European Language Resources Association, Marseille, France. URL: https://aclanthology.org/2022.fnp-1.16.

D. Mariko, H. A. Akl, E. Labidurie, S. Durfort, H. de Mazancourt, and M. El-Haj. 2021. The Financial Document Causality Detection Shared Task (FinCausal 2021). In Proceedings of the 3rd Financial Narrative Processing Workshop, pages 58–60. Association for Computational Linguistics, Lancaster, United Kingdom. URL: https://aclanthology.org/2021.fnp-1.10.

Moreno-Sandoval, Antonio and Carbajo-Coronado, Blanca and Porta-Zamorano, Jordi and Torterolo-Orta, Yanco-amor and Samy, Doaa. 2025. The Financial Document Causality Detection Shared Task (FinCausal 2025). Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal) - at COLING-2025

Annexure

	English Dataset				**Spanish Dataset**			
	XLM-Roberta-Base-Squad2		**XLM-Roberta-Large-Squad2**		**XLM-Roberta-Base-Squad2**		**XLM-Roberta-Large-Squad2**	
	Without Fine Tuning	**Fine Tuned**	**Without Fine Tuning**	**Fine Tuned**	**Without Fine Tuning**	**Fine Tuned**	**Without Fine Tuning**	**Fine Tuned**
Semantic Answer Similarity (SAS)	0.73	0.93	0.78	**0.96**	0.76	**0.96**	0.79	0.96
Exact Match	0.21	0.68	0.28	**0.72**	0.16	**0.76**	0.17	0.71

.

Table 2. Results obtained on the test data for our different models

	Practice Dataset				Development Dataset			
	XLM-Roberta-Base-Squad2		XLM-Roberta-Large-Squad2		XLM-Roberta-Base-Squad2		XLM-Roberta-Large-Squad2	
	Without Fine Tuning	Fine Tuned	Without Fine Tuning	Fine Tuned	Without Fine Tuning	Fine Tuned	Without Fine Tuning	Fine Tuned
(English)								
Semantic Answer Similarity (SAS)	0.82	0.92	0.75	**0.96**	0.80	**0.94**	0.76	0.91
Exact Match	0.44	0.62	0.33	**0.74**	0.34	**0.70**	0.26	0.61
(Spanish)								
Semantic Answer Similarity (SAS)	0.73	0.94	0.66	**0.95**	0.76	0.95	0.76	**0.96**
Exact Match	0.13	0.73	0.17	**0.71**	0.16	**0.82**	0.16	0.72

Table 3. Results obtained on the Practice and development data for our different models