The Financial Document Causality Detection Shared Task (FinCausal 2025)

Antonio Moreno Sandoval[1], Blanca Carbajo Coronado[1], Jordi Porta Zamorano[1], Yanco Amor Torterolo Orta[1], Doaa Samy[2]

[1]Universidad Autónoma de Madrid, Spain
[2]Universidad Complutense de Madrid, Spain

Correspondence: antonio.msandoval@uam.es

Abstract

We present the Financial Document Causality Detection Task (FinCausal 2025), a multilingual challenge designed to identify causal relationships within financial texts. This task comprises English and Spanish subtasks, with datasets compiled from British and Spanish annual reports. Participants were tasked with identifying and generating answers to questions about causes or effects within specific short texts. The dataset combines extractive and generative question-answering (QA) methods, with abstractly formulated questions and directly extracted answers from the text. Systems performance is evaluated using exact matching and semantic similarity metrics. The challenge attracted submissions from 10 teams for the English subtask and 10 teams for the Spanish subtask. FinCausal 2025 is part of the 6th Financial Narrative Processing Workshop (FNP 2025), hosted at COLING 2025 in Abu Dhabi.

Keywords: causal detection, QA task, financial documents, NLP

1 Introduction

Financial analysis relies on factual data to provide a clear view of current conditions, but it also needs deeper insights to understand how and why these facts have come to be. The ultimate goal of FinCausal 2025 is to determine, regarding a given context, which events or chain of events can cause a financial object to be modified or an event to occur.

Historically, extracting cause-effect relationships has been primarily extractive, as demonstrated in previous iterations of the FinCausal task (Mariko et al., 2021; Mariko et al., 2022; Moreno-Sandoval et al., 2023). However, 2025 task is framed as a question-answering task, requiring systems to respond to causality-focused questions, with their answers assessed through exact matching and similarity metrics.

The task comprises two subtasks, one in English and one in Spanish. Participants were required to provide the answer for each question using any method of their choice. Both datasets were created from annual reports, making them suitable for testing of multilingual models.

Annual reports detail a company's economic, financial, and operational performance during the year, including management insights, corporate governance, and social responsibility. For this task, we focus solely on the narrative sections, excluding the financial statements.

2 The dataset

In both subtasks, causality was described as a relationship in which two events are connected, with one event, occurring earlier in time, acting as the trigger for the other. Causes and their effects may be represented by agents or facts. There are two primary types of causes:

1. **Causes justifying a statement**. For example: 'This is my final report since I have been succeeded as President of the Commission as of January 24, 2019.'

2. **Causes explaining a result**. For example: 'In Spain, revenue grew by 10.8% to 224.9 million euros, due to an increase in cement volume accompanied by a more moderate price increase.'

To create the dataset, a question was formulated for each context asking for either the cause or the effect, followed by a corresponding answer. Each context contains a cause-effect relationship, though not every sentence in the sample is case of causality.

A maximum of two questions per context were allowed in cases involving complex causal relationships, such as a chain of three or more elements or non-linear relationships. Contexts lacking a clear or complete causal relationship, or those express-

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 214–221
January 19–20, 2025.

ing conditions, purposes, or concessions, were excluded. This exclusion was based solely on the provided context, without drawing inferences from any external knowledge.

The dataset comprises three key components:

- **Context**: The original short text extracted from financial annual reports.

  In English, the context ranged from 9 to 191 words, with an average of 43 words. In Spanish, it spanned 4 to 255 words, averaging 46 words.

  Each context has its own ID. Sequential IDs were given when two questions were formulated for a single context (with letters XX.a, XX.b, and XX.c, etc.) and when the context was divided into multiple parts (with numbers XX.XX.1, XX.XX.2, XX.XX.3...).

- **Question**: Formulated to identify the other half of a causal relationship, either the cause or the effect. It is abstractive; it does not reproduce the context directly. For example, questions in English may be formulated as follows: 'What triggered X?', 'What was the outcome of X?', or 'What influence did X have on Y?'. Similarly, in Spanish, examples include: '¿Qué originó X?', '¿Cuál es el resultado de X?', or '¿Qué influencia tiene X sobre Y?'. There was an emphasis on not inserting external data or superfluous details.

  Questions were framed in third person or impersonally if the source text used the first person.

- **Answer**: The cause or effect in question, extracted directly from the text without altering the structure. It could be comprised of one or multiple sentences as required semantically. Causal or consecutive connectors were omitted whenever possible, provided that the coherence with the question was maintained.

  When multiple text chains were possible answers, the option with the greatest level of detail was selected. In contexts with two questions, one answer could partially or fully match the other one.

Both the English and Spanish dataset sizes are shown in Table 1. These files are available in UTF-8 plain text and CSV formats, with each line containing four columns separated by ';':

```
ID;Context;Question;Answer
```

Additional information can be obtained at https://www.lllf.uam.es/wordpress/fincausal-25/. The task has been managed through Codalab (https://codalab.lisn.upsaclay.fr/competitions/19936).

2.1 The English subtask

The English dataset was drawn from a corpus on annual reports key sections provided by Lancaster University (El-Haj et al., 2019). This corpus includes reports from both financial and non-financial firms listed on the London Stock Exchange (LSE) Main Market or the Alternative Investment Market (AIM). For this task, we focused on annual reports from 2017. Participants received text block samples from the corpus, each containing at least one causal relationship. The shortest context consisted of 4 words, the longest reached 191 words, with an average of 43 words per fragment. Two examples from the dataset are presented in Table 2.

Set	English	Spanish
Training	1,999	2,000
Test	499	500

Table 1: Datasets.

2.2 The Spanish subtask

The dataset was sourced from a corpus of 305 Spanish financial annual reports from 2014 to 2018, FinT-esp (Moreno Sandoval et al., 2020). Participants were provided with a sample of shorts texts extracted from the corpus, consisting of a paragraph with at least one causal relationship. The longest context contains 255 words, while the average number of words per fragment is 46. Table 3 presents two samples from the dataset.

The 5,000 fragments that make up the entire FinCausal dataset were created by four linguists with expertise in annotation and prompting.

3 Competition: participants and systems

Initially, 41 users registered for the challenge. Of these, 14 submitted at least one entry to the Codalab server, and ultimately 11 different groups participated in the ranking. Among them, 9 groups took part in both the English and Spanish tasks, while 1 group participated only in the English task,

Context	Question	Answer
In October 2016, we announced an implementation agreement to sell ACR to two Shenzhen government sponsored investment companies. This approval process remains ongoing and, as a result, we did not value ACR on an imminent sales basis as at 31 March 2017.	Why was ACR not valued on an imminent sales basis as of March 31, 2017?	This approval process remains ongoing
The Board has resolved that, in view of the size of the Board, it is most appropriate for matters of remuneration to be dealt with by the Board as a whole.	What was the implication of the Board's size?	it is most appropriate for matters of remuneration to be dealt with by the Board as a whole

Table 2: Sample for the English subtask.

Context	Question	Answer
Por otra parte, Banco Sabadell se mantiene como referente financiero del sector público gracias a la innovación en productos y servicios para la administración.	¿A qué se debe que Banco Sabadell se mantenga como referente financiero del sector público?	a la innovación en productos y servicios para la administración
La plantilla aumentó un 2,6% dado que se han puesto en marcha nuevas líneas y que ha aumentado la producción.	¿Qué explica el aumento de la plantilla de un 2,6%?	se han puesto en marcha nuevas líneas y que ha aumentado la producción

Table 3: Sample for the Spanish subtask.

and another group participated only in the Spanish task. Nearly 500 submissions were received during the first 11 days of testing. A wide variety of countries are represented among the final participants: China, Austria, India (x4), Singapore, Denmark, Egypt, and Spain.

4 Evaluation metrics

Semantic Answer Similarity (SAS), as introduced in Risch et al. (2021), is the primary metric used to measure how similar two texts are based on their semantic meaning rather than just word-for-word matching. It is particularly useful in evaluating responses in tasks like abstractive question-answering. SAS utilizes pre-trained language models like BERT (Devlin et al., 2019) or Sentence Transformers (Reimers and Gurevych, 2019) to generate text embeddings and then computes cosine similarity between these embeddings to assess how closely two pieces of text align in meaning, even if they use different words or structures. This allows for more accurate evaluation of content that conveys the same idea but is expressed differently.

We chose to include SAS as a metric because, in FinCausal 2023, the majority of the participating models were generative prompting-based models (based on GPT), and a traditional metric such as Exact Match (EM) alone proved inadequate for accurately evaluating their outputs. For FinCausal 2025, we have used the Paraphrase Multilingual Mpnet Base V2 model[1] using a Sentence Transformer architecture built on a pre-trained XLM-RoBERTa model (Conneau et al., 2020) to give support to the Spanish and English subtasks, converting text into 768-dimensional vectors.

Additionally, we used Exact Match (EM) as a secondary metric. It measures the accuracy by checking whether the model's generated answer matches the reference answer exactly, word by word.

Both metrics, SAS and EM, are averages over the individual values of the examples to which they are applied.

5 Results and discussion

5.1 The baseline

The baseline for the competition was conceived as a minimal starting point to serve as a reference, while also testing the dataset. In order to achieve this, a basic extractive QA pipeline was selected to satisfy the EM metric and produce scores for the SAS metric. The Transformers library (Wolf et al., 2020) was utilized for both English and Spanish tasks, employing the generic

[1] sentence-transformers/paraphrase-multilingual-mpnet-base-v2

model class `AutoModelForQuestionAnswering` with `from_pretrained()`. In both cases, the datasets were converted into the SQuAD dataset format (Rajpurkar et al., 2018) to simplify preprocessing. The following is an example of this format: `{'id': '', 'context': '', 'question': '', 'answers': {'text': []}, 'answer_start': []}`. The key change is the inclusion of the position of the answer within the context, represented by an index in the `answer_start` field.

The training process was straightforward, applying default hyperparameters. The HuggingFace Trainer was used with the default data collator. For the English task, the model `distilbert/distilbert-base-uncased` (Sanh et al., 2019) was employed without further experimentation, as the scores were deemed sufficient for the baseline. Conversely, the Spanish task required some additional experimentation to achieve comparable results, ultimately selecting `PlanTL-GOB-ES/roberta-large-bne` (Fandiño et al., 2022) as the counterpart to the English model. The English baseline scores were 0.7373 for SAS and 0.3286 for EM, while the Spanish baseline reached 0.7244 for SAS and 0.2515 for EM.

5.2 English task

Ten teams, in addition to the baseline system, participated in the English subtask. All of these teams, except for Sarang, also competed in the Spanish subtask.

Team Nirvanatear (Jonathan Zhou) employed a fine-tuned large language model (LLM) approach. Specifically, he fine-tuned LLMs (gpt4o-mini, Llama 3.1-8B) on causality QA data to directly answer test questions through simple prompting. The team conducted extensive experimentation, varying LLMs, prompt configurations, data selection (language-specific, bilingual, or validation-based), and the inclusion of additional user-generated QA data. Ensemble methods were also explored. Their English task submission utilized a gpt-4o-mini model fine-tuned on a bilingual dataset, prompted with: 'You are a helpful assistant. Read the paragraph and succinctly answer the question about causality that follows.'

The **TU Graz Data team** adopted the same architecture for both tasks. They trained Llama 3.1 8B and 70B models using LoRA-based fine-tuning and a few-shot optimized prompt. A bilingual dataset was used, alternating between Spanish and English lines to train multilingual models. Model outputs were compared using cosine similarity, with GPT-4 serving as a tiebreaker.

Team Sarang, from NIT Trichy, employed a simpler approach without external databases. Their system involved selecting `consciousAI/question-answering-roberta-base-s`. They refined the FinCausal-2025 development set by filtering it to include only rows where the answer appeared as a substring of the context. The preprocessed dataset was then split into a 90:10 ratio for training and validation. Following this, the selected checkpoint was fine-tuned to enhance performance. Finally, the team leveraged the capabilities of `Gemma-2-9B` through prompt engineering to improve results further.

Team OraGenAI, from Oracle, India, introduced the Knowledge Utilization Framework (KULFi), a novel approach to enhance LLM reasoning capabilities in financial causal reasoning. KULFi addresses the limitations of human-guided prompt engineering and computationally intensive fine-tuning by automating prompt optimization through Teacher-Student interactions. Key components of KULFi include:

- Auto CoT transfer: The Teacher LLM generates reasoning chains (Chain of Thought) to guide the Student LLM.
- Auto task alignment: The Teacher provides task-specific instructions, iteratively refining the Student's performance.

Laith Team employs the XLM-RoBERTa-large model, a multilingual transformer, to perform extractive question answering (QA) tasks. The model has been fine-tuned on both English and Spanish datasets. This bilingual approach equips the model with the capacity to generalize across languages, a crucial attribute for the multilingual nature of FinCausal tasks.

The training process involved parameter tuning, with a batch size of 16 for both training and evaluation. A learning rate of 2e-5, coupled with a weight decay of 0.01, was employed to optimize the model's learning trajectory. The model was trained for 10 epochs, with evaluation conducted at the conclusion of each epoch to monitor its progress.

To ensure efficient processing of multilingual text, they leveraged the XLM-RoBERTaTokenizerFast for tokenization. This tokenizer effectively handles multilingual subword tokenization, enabling the model to process text

from diverse languages. To accommodate longer contexts, inputs were tokenized with a maximum sequence length of 384 tokens and a stride of 128, allowing for overlapping windows to capture comprehensive information.

The system employs a traditional extractive framework, enhanced with multilingual capabilities through training on both English and Spanish datasets. This allows the model to directly identify relevant text spans from the input document to answer questions. The model's ability to generalize across languages makes it well-suited for multilingual FinCausal tasks.

CLRG Team submitted the results achieved with XML-RoBERTa base and large models fine-tuned for Extractive QA on various languages using the SQuAD dataset (Rajpurkar et al., 2016) and tuned with FinCausal 2025 data for each sub-task.

The remaining teams did not provide detailed system descriptions.

5.3 Spanish task

Team Nirvanatear, The TU Graz Data team, Team Sarang, LaithTeam and Team OraGenAI employed the same systems outlined in the English subtask (Section 5.2) to compete in the Spanish subtask.

Team LenguajeNatural.AI employs the Supernova generative model, a private model based on a combination of publicly available multilingual models ranging from 7B to 8B parameters, which was pre-trained used a corpus of supervised tasks for Spanish and fine-tuned on a variety of Spanish intruction-following datasets. The model was then fine-tuned with QLoRA with the FinCausal the training set. At inference time, they use a fuzzy match algorithm to ground predicted answers in the context information of the question.

In general, all teams that participated in both subtasks performed slightly better in Spanish. The reason for this can only be found by analysing each team's results in detail. In the following sections we provide some examples.

5.4 Taxonomy of participant systems

Table 5 compares the systems that were described by the participants. There is a wide variety of approaches; however, in general terms, participants tended to favor generative models. Fine-tuning was also a commonly preferred option.

5.5 Error analysis

The errors in the teams' predictions, both in English (see Table 7) and Spanish (see Table 8), stem primarily from two issues. First, purpose-based relationships are often confused with cause-effect relationships. This happens when a response describing a goal or desired outcome is mistakenly presented as the cause of an event. Additionally, in some cases, elements from purpose-based or even concessive relationships (*although*, *despite*...) are added to the correct response, introducing unnecessary contextual information that is irrelevant to answering the question. This type of error is particularly common in cases where SAS scores are high, but EM is 0.

Second, errors with lower SAS scores are typically the result of minimal overlap between the generated response and the expected one. In such cases, the models fail to properly identify the key elements of the causal relationship or exhibit poor understanding of the question's context.

6 Conclusions

After several editions dedicated to the extraction of cause-effect segments in financial annual reports, FinCausal 2025 has been approached as a QA task. The challenge includes both English and Spanish subtasks, each supported by datasets containing 2,500 samples. This year's edition incorporated the SAS metric alongside the EM metric for a more comprehensive evaluation of participants' responses. In fact, the SAS metric was suggested by participants of the previous FinCausal 2023.

In the English subtask, Team Nirvanatear achieved top performance by fine-tuning gpt4o-mini on targeted datasets, while the TU Graz Data Team employed multilingual models with LoRA-based fine-tuning and bilingual datasets. Team Sarang showcased the potential of lightweight approaches without external databases. The Laith system employs a traditional extractive framework based on the multilingual XLM-RoBERTa-large model. The model has been fine-tuned on both English and Spanish FinCausal datasets, without external databases. OraGenAI introduced KULFi, a framework automating prompt optimization through teacher-student interactions. Many teams also used these systems in the Spanish subtask, demonstrating the adaptability of their models. Notably, Team LenguajeNatural.AI highlighted the importance of language-specific resources.

Ranking	Team	SAS	Exact Match
1	Team nirvanatear (Jonathan Zhou, China)	0.9779 (1)	0.8798 (1)
2	TU Graz Data Team (Graz University of Technology, Austria)	0.9732 (2)	0.8637 (2)
3	Sarang (National Institute of Technology ,Trichy, India)	0.9674 (3)	0.7014 (7)
4	CLRG (n/a)	0.9604 (4)	0.7214 (6)
5	Semantists (Institute for Infocomm Research, Singapore)	0.9598 (5)	0.7435 (5)
5	LaithTeam (Copenhagen University, Denmark)	0.9598 (5)	0.7615 (4)
7	CUFE (Cairo University, Egypt)	0.9595 (7)	0.8277 (3)
8	OraGenAIOrganisation (Oracle, India)	0.9244 (8)	0.3527 (9)
9	RGIPT (India)	0.9086 (9)	0.5110 (8)
10	PresiUniv (Dpt. CSE, Presidency Univ, Bangalore, India)	0.8241 (10)	0.2244 (11)
11	Baseline (LLI-UAM, Spain)	0.7373 (11)	0.3287 (10)

Table 4: English results

Team	Discriminative	Generative	Fine-tuning	Prompting	Quantization
Team Nirvanatear	✗	✓	✓	Simple	✗
OraGenAIOrganisation	✗	✓	✗	CoT	✗
Al Laith	✓	✗	✓	✗	✗
Sarang	✗	✓	✓	Simple	✓
RGIPT	✗	✓	✗	CoT+FS/FS	✗
TU Graz	✗	✓	✓	✗	✓
PresiUniv	✓	✗	✗	✗	✗
LenguajeNatural.AI	✗	✓	✓	Simple	✓
CLRG	✓	✗	✓	✗	✗

Table 5: Systems comparison. In Prompting, Simple means a simple prompt or instruction, CoT stands for Chain of Thoughts and FS stands for Few Shot.

Ranking	Team	SAS	Exact Match
1	TU Graz Data Team (Graz University of Technology, Austria)	0.9841 (1)	0.8703 (2)
2	Team nirvanatear (Jonathan Zhou, China)	0.9801 (2)	0.8782 (1)
3	LenguajeNatural.AI (Spain)	0.9787 (3)	0.8164 (4)
4	LaithTeam (Copenhagen University, Denmark)	0.9756 (4)	0.8084 (5)
5	CUFE (Cairo University, Egypt)	0.9755 (5)	0.8224 (3)
6	CLRG (n/a)	0.9607 (6)	0.7166 (7)
7	Semantists (Institute for Infocomm Research, Singapore)	0.9555 (7)	0.7525 (6)
8	OraGenAIOrganisation (Oracle, India)	0.9219 (8)	0.0898 (9)
9	RGIPT (India)	0.8987 (9)	0.0619 (10)
10	PresiUniv (Dpt. CSE, Presidency Univ, Bangalore, India)	0.7520 (10)	0.0140 (11)
11	Baseline (LLI-UAM, Spain)	0.7244 (11)	0.2515 (8)

Table 6: Spanish results

Errors primarily stemmed from confusing causal relationships with purpose-based statements or introducing irrelevant context, such as concessive phrases. While semantic similarity scores were high, lower exact match scores indicated challenges in extracting precise causal elements.

The 2025 edition surpassed the performance of FinCausal 2023 (Moreno-Sandoval et al., 2023),

Context	Question	Answer	Result	SAS	Exact match
In accordance with the Company's stated dividend policy, the Board recommends a further quarterly dividend of 3.57p per Ordinary Share, payable on 30 April 2018 to shareholders on the register on 6 April 2018. Total dividends paid for the year therefore amount to 14.04p per Ordinary Share equivalent to a dividend yield of 4.1 per cent at the year-end.	Why does the total dividends paid for the year amount to 14.04p per Ordinary Share, equivalent to a dividend yield of 4.1 per cent at the year-end?	the Board recommends a further quarterly dividend of 3.57p per Ordinary Share, payable on 30 April 2018 to shareholders on the register on 6 April 2018	In accordance with the Company's stated dividend policy, the Board recommends a further quarterly dividend of 3.57p per Ordinary Share, payable on 30 April 2018 to shareholders on the register on 6 April 2018	0.980	0
Deloitte LLP has been the Company's external auditor since launch in 2010, and this is its eighth consecutive annual audit. As a result of its work during the year, the Audit Committee concluded that Deloitte acted in accordance with its terms of reference.	What were the consequences of Deloitte LLP being the Company's external auditor for eight consecutive annual audits?	the Audit Committee concluded that Deloitte acted in accordance with its terms of reference	its work during the year, the Audit Committee concluded that Deloitte acted in accordance with its terms of reference	0.978	0
Share based charges increased by £0.7m due to the continued investment in the Franchise Incentive Plan and management share options to ensure both Franchisees and management are aligned with the Group's objectives and rewarded based on the performance of the Group.	What motivated the increase in share-based charges by £0.7m?	the continued investment in the Franchise Incentive Plan and management share options	the continued investment in the Franchise Incentive Plan and management share options to ensure both Franchisees and management are aligned with the Group's objectives and rewarded based on the performance of the Group	0.883	0
Communication is key to innovation in our business. Breaking down silos and sharing best practice allows us to leverage the expertise in our business and provide the best service to our customers. Because of this, DS Smith invested in enhancing our communication and collaboration platforms	What factor led DS Smith to invest in enhancing their communication and collaboration platforms?	Communication is key to innovation in our business. Breaking down silos and sharing best practice allows us to leverage the expertise in our business and provide the best service to our customers	Breaking down silos and sharing best practice allows us to leverage the expertise in our business and provide the best service to our customers	0.752	0

Table 7: Examples of errors in English.

Context	Question	Answer	Result	SAS	Exact match
En este contexto, GRIDSOL representa un gran impulso para integrar fuentes de energía renovables gracias a la generación flexible. Demostrando la adecuación de los Smart Renewable Hubs para redes continentales e insulares con el fin de lograr un sistema de energía más seguro y limpio.	¿Qué supone la generación flexible?	GRIDSOL representa un gran impulso para integrar fuentes de energía renovables	En este contexto, GRIDSOL representa un gran impulso para integrar fuentes de energía renovables	0.985	0
En este caso, el impacto directo recogido en las cuentas de 2017 se ha estimado en 2,6 millones de euros, concentrado en los costes del basmati (que afecta especialmente al mercado europeo) ya que la variación de otras variedades de fragante se produjo al final de año con un nivel de alerta superior y, en todo caso, será objeto de las negociaciones con la distribución en 2018.	¿Por qué el impacto directo recogido en las cuentas de 2017 se ha estimado en 2,6 millones de euros, concentrado en los costes del basmati?	la variación de otras variedades de fragante se produjo al final de año con un nivel de alerta superior	la variación de otras variedades de fragante se produjo al final de año con un nivel de alerta superior y, en todo caso, será objeto de las negociaciones con la distribución en 2018	0.802	0
La orientación al cliente nos impulsa a trabajar en la gestión de calidad de nuestras autopistas	¿Cuál es la razón de que trabajen en la gestión de calidad de sus autopistas?	La orientación al cliente	La orientación al cliente nos impulsa	0.880	0
Storstockholms Lokaltrafik AB, empresa responsable de la red de transportes de Estocolmo, ha firmado dos ampliaciones durante el pasado año, adquiriendo 20 nuevos tranvías: 10 de cuatro módulos y otros 10 de tres módulos, con lo que dispondrá de 42 tranvías Urbos en su flota para la capital sueca.	¿Por qué se podrá disponer de 42 tranvías Urbos en su flota para la capital sueca?	Storstockholms Lokaltrafik AB, empresa responsable de la red de transportes de Estocolmo, ha firmado dos ampliaciones durante el pasado año, adquiriendo 20 nuevos tranvías: 10 de cuatro módulos y otros 10 de tres módulos	adquiriendo 20 nuevos tranvías: 10 de cuatro módulos y otros 10 de tres módulos, con lo que dispondrá de 42 tranvías Urbos en su flota para la capital sueca	0.781	0

Table 8: Examples of errors in Spanish.

even with the paradigm shift from an extractive to a question-answering approach. The doubling of participating teams underscores the growing interest and rapid advancement of generative AI-based technologies.

Acknowledgments

We thank our dedicated annotator, Paula Gozalo, who contributed to creating the datasets. This publication is part of the project GRESEL (PID2023-151280OB-C21) funded by the Spanish Ministry of Science and Innovation and Universities.

We also gratefully acknowledge the financial support received by the second author through a FPU grant (FPU20/04007) awarded by the Spanish Ministry of Science, Innovation and Universities.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Mahmoud El-Haj, Steven Young, and Paul Rayson. 2019. Annual reports key sections corpora 2003 to 2017 [dataset].

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (FinCausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 105–107, Marseille, France. European Language Resources Association.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2021. The financial document causality detection shared task (FinCausal 2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 58–60, Lancaster, United Kingdom. Association for Computational Linguistics.

Antonio Moreno Sandoval, Ana Gisbert, and Helena Montoro. 2020. FinT-esp: a corpus of financial reports in Spanish. In Miguel Fuster-Márquez, Carmen Gregori-Signes, and José Santaemilia Ruiz, editors, *Multiperspectives in Analysis and Corpus Design*, pages 89–102. Editorial Comares, Granada, Spain.

Antonio Moreno-Sandoval, Jordi Porta-Zamorano, Blanca Carbajo-Coronado, Doaa Samy, Dominique Mariko, and Mahmoud El-Haj. 2023. The financial document causality detection shared task (FinCausal 2023). In *2023 IEEE International Conference on Big Data (BigData)*, pages 2855–2860.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic Answer Similarity for Evaluating Question Answering Models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.