UPB at FinCausal-2020, Tasks 1 & 2: Causality Analysis in Financial Documents using Pretrained Language Models

Marius Ionescu[1], Andrei-Marius Avram[1,2], George-Andrei Dima[1,3], Dumitru-Clementin Cercel[1], Mihai Dascalu[1]
University Politehnica of Bucharest[1]
Research Institute for Artificial Intelligence, Romanian Academy[2]
Military Technical Academy Ferdinand I[3]
{ionescumarius23, avram.andreimarius}@gmail.com,
andrei.dima@mta.ro, {dumitru.cercel, mihai.dascalu}@upb.ro

Abstract

Financial causality detection is centered on identifying connections between different assets from financial news in order to improve trading strategies. FinCausal 2020 - Causality Identification in Financial Documents – is a competition targeting to boost results in financial causality by obtaining an explanation of how different individual events or chain of events interact and generate subsequent events in a financial environment. The competition is divided into two tasks: (a) a binary classification task for determining whether sentences are causal or not, and (b) a sequence labeling task aimed at identifying elements related to cause and effect. Various Transformer-based language models were fine-tuned for the first task and we obtained the second place in the competition with an F1-score of 97.55% using an ensemble of five such language models. Subsequently, a BERT model was fine-tuned for the second task and a Conditional Random Field model was used on top of the generated language features; the system managed to identify the cause and effect relationships with an F1-score of 73.10%. We open-sourced the code and made it available at: https://github.com/avramandrei/FinCausal2020.

1 Introduction

Financial news contain various descriptions of causes and effects between financial objects that influence sales, employment, earnings, or stock prices. Causal analysis can be used to detect correlation between news and prices of different assets (Qu and Kazakov, 2019) by identifying different consequence, that in return can lead to major events as a financial crisis (Stavroglou et al., 2017; Tiffin, 2019) or possible arbitrage opportunities (Stavroglou et al., 2017). For example, given the follow-up statement regarding Brexit *"The Leave win led to an 11 percent drop in GBP/USD overnight."*, we can infer that *"The Leave win"* is the cause for *"an 11 percent drop in GBP/USD overnight"*, which represents the effect and indicates a change in price of the GBP/USD currency pair (Mariko et al., 2020).

Nowadays, the high volume of published news and financial documents hinders and renders unfeasible the manual analysis of all articles in order to extract implications on the economy. Thus, a need for developing new tools for causality extraction emerged to facilitate this process. Methods based on pattern detection (Khoo et al., 2000), on combinations of patterns and rules (Sorgente et al., 2013), and on attention mechanisms with different embeddings (Li et al., 2019) were introduced in the field of causal detection. A shared task was proposed at the Financial Narrative Processing Workshops, namely FinCausal-2020 (Mariko et al., 2020), to evaluate the performance of each participant system on two tasks defined on a financial corpus proposed by YseopLab. The *first task* consisted in building a classifier to identify whether a sentence contains a financial causality or not, whereas the *second task* extracted causes and also their corresponding effects from a given text.

The rest of the paper is organized as follows. The second section introduces our proposed methods for tackling the two sub-tasks, while the third section presents our results, together with experimental setup and an error analysis. The paper ends with conclusions and proposals of future work.

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 55–59
Barcelona, Spain (Online), December 12, 2020.

2 Dataset

The organizers made available 3 datasets for both tasks: trial, practical, and evaluation. The trial dataset for the first task contained 8580 samples, out of which only 569 (i.e., 6,69% from total samples) included causal events. The practice dataset has approximately the same distribution as trial, with 13478 samples out of which 1010 were labeled as causal (i.e., 7,49% from total samples). The evaluation dataset was kept blind, but it has the same distribution as trial and practice.

The dataset for the second taks largely corresponds to samples from Task 1 that are labelled as causal, with the specification of causal and effect sub-strings. The trial dataset contains 641 samples, while the practical dataset contains 1109 samples. The evaluation dataset contained 638 samples with removed annotations, i.e., the text is revealed, but there are no causal or effect representations.

3 Proposed Solutions

Five Transformer-based language models (Vaswani et al., 2017) were considered for the *the first task* and were fine-tuned on the FinCausal-2020 dataset, namely: BERT (base and large) (Devlin et al., 2018), ALBERT (base and large) (Lan et al., 2019), RoBERTA (base and large) (Liu et al., 2019), SciBERT (base) (Beltagy et al., 2019), and FinBERT (base) (Araci, 2019). The fine-tuning mechanism was the one proposed by Devlin et al. (2018) who take the embedding of the first token and project it into a scalar that represents a probability. Then, we minimize the Kullback–Leibler divergence between the distribution of the dataset and the distribution predicted by the model by using the Adam optimizer (Kingma and Ba, 2014). We also experimented with various classifier ensembles taking into account different combinations of the fine-tuned language models as these ensembles surpass in general the performance of individual models; majority voting was considered for the final labeling of each sentence.

The *second task* was operationalised as a token labeling, similar to part-of-speech tagging (Bohnet et al., 2018) or named entity recognition (Dumitrescu and Avram, 2019), by marking the corresponding tokens of the causes and effects sequences with *CAUSE* and *EFFECT* labels, whereas the rest of the tokens were marked as *O*. Inside-Outside-Beginning (IOB) format was used for the labels to support the identification of the sequences at inference time. BERT-base models were used to obtain contextualized embeddings for each token, and then a Conditional Random Field (CRF) (Lafferty et al., 2001) was trained to predict the most probable sequence of labels for a given input text.

Although the IOB format and the CRF aided in modeling the problem as a token labeling task, the extraction of causes and effects is harder at inference because the model can make an incorrect prediction in the middle of a cause/effect sequence, or predict a cause/effect token in the middle of *O* tags. Several heuristics were introduced to mitigate this issue when extracting the causes and effects, namely:

- If a cause or effect sequence has a length lower than 4, it is ignored.
- In case that the model does not predict a cause or an effect in a sequence, all the remaining *O* tags become the missing cause or effect.
- If a different predicted sequence (e.g. "EFFECT EFFECT CAUSE CAUSE EFFECT EFFECT") is found inside the initial sequence, the inside predicted sequence has a length lower than 4, and the other part of the initial sequence does not have the beginning tag[1], then the inside predicted sequence is ignored and the two initial sequences are combined.
- After the tokens are labeled as either *CAUSE* or *EFFECT*, the original text is retrieved by identifying the boundary words using the Levenshtein distance (Levenshtein, 1966); a word is considered a boundary word if it has a Levenshtein distance equal to 0 and a length higher than 3. Afterwards, all words between boundary words are labeled as either *CAUSE* or *EFFECT*.

[1] The beginning tag can be either B-EFFECT or B-CAUSE.

	Dev Set			Test Set		
Model	Precision	Recall	F1 Score	Precision	Recall	F1 Score
ALBERT-Base	95.71	95.88	95.78	96.75	96.76	96.75
SciBERT-Base	95.67	95.99	95.75	96.77	96.83	96.80
FinBERT-Base	93.88	94.71	93.92	94.08	94.30	94.18
BERT-Large	**95.81**	**96.15**	**95.77**	97.10	97.02	97.05
RoBERTa-Large	95.69	95.29	95.46	97.35	97.30	97.32
Ensemble Model	-	-	-	**97.53**	**97.59**	**97.55**
Baseline	95.26	95.21	95.23	-	-	-
Winning Team	-	-	-	97.73	97.76	97.74

Table 1: Task 1 performance of our methods against the baseline and the winning team solution.

	Dev Set			Test Set		
Model	Precision	Recall	F1 Score	Precision	Recall	F1 Score
BERT-CRF	66.66	74.34	68.85	75.61	72.13	73.10
Baseline	50.98	51.74	51.06	-	-	-
Winning System	-	-	-	94.78	94.70	94.71

Table 2: Task 2 performance of our method against the baseline and the winning team solution.

4 Evaluation

4.1 Experimental Setup

We fine-tune five pretrained models provided by Hugging Face[2]: BERT, RoBERTa, ALBERT, SciBERT, and FinBERT. The models are trained using a Tesla P100-PCIE-16GB GPU. Grid search is performed to determine best hyper-parameters, learning rate, and maximal length of sequence. We use a maximal length of the sequence of 256 for *base* models and 384 for *large* models. Learning rate is set to 2×10^{-6}, with the exception of ALBERT-Base where a learning rate of 2×10^{-5} is preferred. Tests before the evaluation submission were made on a dataset that contained data from trial and practice, separated in 80% train and 20 % validation partitions, while keeping the same distribution as previous datasets.

4.2 Results

Table 1 introduces the results for the first task obtained using the best ensemble of language models, as well as each individual language model, on both the validation and the evaluation sets. The best ensemble contained the following models: ALBERT-Base, SciBERT-Base, BERT-Large, FinBERT-Base and RoBERTA-Large, and it outperformed RoBERTa-Large with 0.23% on the evaluation set, obtaining an 97.55% F1-score, while being marginally behind by 0.19% from the winning model. While considering individual models, BERT-Large obtained the highest score on the validation set with an 95.77% F1-score, surpassing the baseline of the organizers by 0.54%; RoBERTA-Large obtained the highest score on the evaluation set with an 97.32% F1-score.

The results on the second task using BERT-Base and CRF are depicted in Table 2. The model outperformed the baseline by 17.79% in terms of F1-score, but it was surpassed by 21.61% on the evaluation set.

4.3 Error Analysis

Several problems were identified when inspecting the trial and practice datasets, and the errors performed by our models. We observed that false positive examples on the first task exhibit statements containing actions or facts, but these actions or facts do not affect other entities in that sentence. For example, a penalty in the first entry from Table 3 may have effects on individuals with debts, but this is not explicitly mentioned in the sentence; thus, the context is not complete to consider it a cause-effect situation. In

[2] https://huggingface.co/transformers/

Index	Text	Pred Label	True Label
0334.00010	The current penalty is 2% per month of the amount of unpaid taxes.	True	False
0360.00009	In the last 90 days, insiders have purchased 817 shares of company stock worth $39,799.	True	False
0588.00013	The average ticket price on TickPick, an online exchange for reselling tickets, plunged 50 percent from $341.19 in 2018 to $170.60 in 2019. With multiple theme park attractions, a film studio tour, and a new exhibit opening in Manhattan, there are other ways for fans of the franchise to consume its content.	False	True
0114.00014	The company had revenue of $64.68 million during the quarter, compared to analyst estimates of $64.51 million. During the same period in the previous year, the business earned $0.73 earnings per share. The firm's quarterly revenue was down 3.9% compared to the same quarter last year.	False	True

Table 3: Misclassified examples for Task1.

the second example, shares purchased by insiders usually have impact on price, but these details are not present. Thus, our models make incorrect predictions because part of these structures are more common in causal statements. In addition, we identified issues in the case of false negative when the cause and the effect are placed in different environments or fields. For example, the cause in the third example is related to attractions located in Manhattan, but the effect is the decrease on an online ticket reselling platform. This is a long dependency and need more specialized knowledge is required to infer these effects. The same situation is encountered in the fourth example, where a company's revenue in this quarter is associate with a 3.9% decrease from same period of last year.

Errors on the second task consist mostly in the incorrect detection of complete sequences of cause or effect. As presented in Table 4, the beginning of our sequences is well identified in most cases, but too few words are considered. This situation may occur from incorrect predictions made by our model or a bad translation from token to text given the post-processing.

Index	Text	Predicted Cause	True Cause	Predicted Effect	True Effect
0332.00009	Net operating profit in the global markets division plunged 26 per cent to 251 billion yen (S$3.19 billion) last year, largely because of a big fall in Europe, and the bank said it struggled in customer business due to sluggish markets.	a big fall in	a big fall in Europe, and the bank said it struggled in customer business due to sluggish markets.	Net operating profit in the global markets division plunged 26 per cent to 251 billion yen	Net operating profit in the global markets division plunged 26 per cent to 251 billion yen (S$3.19 billion) last year
0576.00008	He used the money to help buy a house for his daughter. Now, 229 members of the pension schemes are concerned they may not see their money again.	He used the money to help buy a house for his daughter.	He used the money to help buy a house for his daughter.	Now, 229 members of the pension schemes are concerned they may not see their money	Now, 229 members of the pension schemes are concerned they may not see their money again.

Table 4: Misclassified examples for Task 2.

5 Conclusion and Future Work

As the financial volume of data grows, it becomes harder and harder to analyze chain of events, with corresponding cause and effects. Our solution for the first task at FinCausal-2020 was an ensemble of five Transformer-based language models that placed second on the evaluation leader-board with an 97.55% F1-score. The second task considered a BERT-base model to extract contextualized embeddings for each token, that were further used to train a CRF; an F1-score of 73.10% was achieved on the evaluation set. Possible directions of research include using more powerful language models like XLNet (Yang et al., 2019) or exploring if an additional bidirectional long-short term memory (BiLSTM) on top of the language model to improve the results on either task. In addition, we envision the consideration of specific discourse markers, as well as transfer learning using similar tasks - for example, a model trained on the SNLI corpus from Stanford (Bowman et al., 2015).

References

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

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676.*

Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic tagging with a meta-bilstm model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Stefan Daniel Dumitrescu and Andrei-Marius Avram. 2019. Introducing ronec–the romanian named entity corpus. *arXiv preprint arXiv:1909.01247.*

Christopher SG Khoo, Syin Chan, and Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pages 336–343.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942.*

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2019. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *arXiv preprint arXiv:1904.07629.*

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2020. The Financial Document Causality Detection Shared Task (FinCausal 2020). In *The 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (FNP-FNS 2020)*, Barcelona, Spain.

Haizhou Qu and Dimitar Kazakov. 2019. Detecting causal links between financial news and stocks. In *2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 1–8. IEEE.

Antonio Sorgente, Giuseppe Vettigli, and Francesco Mele. 2013. Automatic extraction of cause-effect relations in natural language text. *DART@ AI* IA*, 2013:37–48.

Stavros K Stavroglou, Athanasios A Pantelous, Kimmo Soramaki, and Konstantin Zuev. 2017. Causality networks of financial assets. *Journal of Network Theory in Finance*, 3(2):17–67.

Mr Andrew J Tiffin. 2019. *Machine learning and causality: the impact of financial crises on growth.* International Monetary Fund.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.