FiNLP at FinCausal 2020 Task 1: Mixture of BERTs for Causal Sentence Identification in Financial Texts

Sarthak Gupta
Munich Re
Munich, Bavaria, Germany
sgupta2@munichre.com

Abstract

This paper describes our system developed for the sub-task 1 of the FinCausal shared task in the FNP-FNS workshop held in conjunction with COLING-2020. The system classifies whether a financial news text segment contains causality or not. To address this task, we fine-tune and ensemble the generic and domain-specific BERT language models pre-trained on financial text corpora. The task data is highly imbalanced with the majority *non-causal* class; therefore, we train the models using strategies such as under-sampling, cost-sensitive learning, and data augmentation. Our best system achieves a weighted F1-score of 96.98 securing 4^{th} position on the evaluation leaderboard. The code is available at https://github.com/sarthakTUM/fincausal

1 Introduction

A causal relation involves two events e_1 and e_2 where one, referred to as 'Cause', is responsible for triggering the other, referred to as the 'Effect'. The causation can be expressed implicitly or explicitly by causal verbs, punctuation, conjunctions, prepositions, or any other linguistic cue that establishes a trigger between the cause and the effect. Explicit causation in a text segment can usually be identified by the presence of causative keywords or characteristic grammatical patterns whereas implicit causality is comparatively harder to detect, usually only through the context. Many studies on Causality Extraction in the text are motivated by (Khoo et al., 1998) in which they identify five different constructs to express explicit cause-effect relation. (Girju, 2003) extended the third construct by identifying lexical syntactic patterns including Noun-Phrases (NP) such as $\langle NP_1$, cause-verb, $NP_2\rangle$ to find causality in the WordNet[1] definitions. (Kim et al., 2007) extends the first and second construct by introducing four categories of the cue phrases for the causality expressions to build an automatic causality extraction framework. More recently, (Dasgupta et al., 2018) proposed an LSTM-CRF architecture for labeling cause and effect tokens in a text segment.

Many prior works on Causality Extraction primarily use the data labeled using weak supervision techniques due to scarcity of the human-labeled data. (Kyriakakis et al., 2019) use Transfer Learning to tackle this issue. To this end, (Mariko et al., 2020) introduced a human-labeled data for causality extraction in financial news text crawled by QWAM.[2]

Shared Task: This shared task consists of two sub-tasks: (i) Binary classification of text segment as *'Causal (C)'* if the text contains causality or *'Non-Causal (NC)'* otherwise, and (ii) Extracting spans of 'cause' and 'effect' within the text segment if it is labelled 'Causal'. The 'effect' can only be a quantified fact, whereas a 'cause' can be a fact or a quantified fact. We propose a system for the first sub-task.

Our Contributions: We augment the data with publicly available Relation Classification and Causality Extraction corpora to increase the number of training instances and leverage Transfer Learning by fine-tuning generic and financial BERT language models (Devlin et al., 2019) on the task data. We train

[1] https://wordnet.princeton.edu/

[2] https://www.qwamci.com/

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 74–79
Barcelona, Spain (Online), December 12, 2020.

individual models using various strategies to tackle the imbalanced target distribution and propose an ensemble using a second-order classifier. Our best system achieves a weighted F1-score of 96.98 securing 4^{th} position on the evaluation leaderboard out of 14 teams.

2 System Description

2.1 Data

The data consists of English text segments from financial news articles released in three installments namely *Trial*, *Practice*, and *Evaluation*. We combine the *Trial* and the *Practice* data to train, test and fine-tune the models. The results on the leaderboard are obtained on the *Evaluation* data. As shown in table 1, the dataset is imbalanced with approximately 92% Non-Causal class. To increase the number of the training instances, we augment the task data with the publicly available (i) training instances labelled with 'Cause-Effect' Relation in the SemEval 2010 Task-8 Relation Classification Corpus (Hendrickx et al., 2010) (ii) training instances labeled with 'Drug-AE' (Drug-Adverse Effect) relation in the ADE corpus (Gurulingappa et al., 2012), and (iii) training instances in Biomedical causal detection dataset provided by (Kyriakakis et al., 2019).

Corpus	**C**	**NC**	**Example (Causal)**
Task (Trial+Practice)	1579	20479	*"Investors are catching on to B. Riley's growth story, driving up shares more than 70% year-to-date."*
SemEval-2010-Task-8	1331	9386	*"The burst has been caused by water hammer pressure."*
Adverse Drug Effect	4271	16625	*"Pirmenol hydrochloride-induced QT prolongation and T wave inversion on electrocardiogram during treatment for symptomatic atrial fibrillation."*
Causaly	1113	887	*"The obstacle of getting older men to undergo circumcision may also be associated with working schedules that may disclose one's circumcision status."*

Table 1: Number of Causal Instances vs. Non-Causal instances along with examples in the corpus

2.2 Models

Modeling the long-range context is a challenge for the LSTM cell (Hochreiter and Schmidhuber, 1997); Therefore the Transformer architecture based BERT (Devlin et al., 2019) along with its proven superiority for Text Classification[3] is an appropriate choice for this binary classification task containing long texts. We explored 4 different pre-trained BERT language models for this task: (i) generic `BERT-base-uncased` provided by (Wolf et al., 2019)[4], (ii) `FinBert-SEC`[5] (Desola et al., 2019) referred to as `FinBert-Combo` in the original publication trained on 10-K SEC filings, (iii) `FinBert-TRC2`[6] (Araci, 2019) trained on Reuters TRC2 corpus, and (iv) `FinBert-FinCom`[7] (Yang et al., 2020) referred to as `FinBERT-FinVocab-Uncased` in the original publication trained on financial communication text such as 10-K & 10-Q corporate reports, earning calls, and analyst reports. In all the models, we replace the `[CLS]` token representation with the mean-pooled embedding of the last layer tokens as shown in figure 1 (right).

We observed that the individual models specialize in different aspects of the confusion matrix, i.e., some models produce more true positives and others produce less false positives primarily due to the difference in the pre-trained weights and the training strategies. Therefore, we ensemble the individual

[3]https://gluebenchmark.com/leaderboard
[4]https://github.com/ThilinaRajapakse/simpletransformers
[5]https://github.com/psnonis/FinBERT
[6]https://github.com/ProsusAI/finBERT
[7]https://github.com/yya518/FinBERT

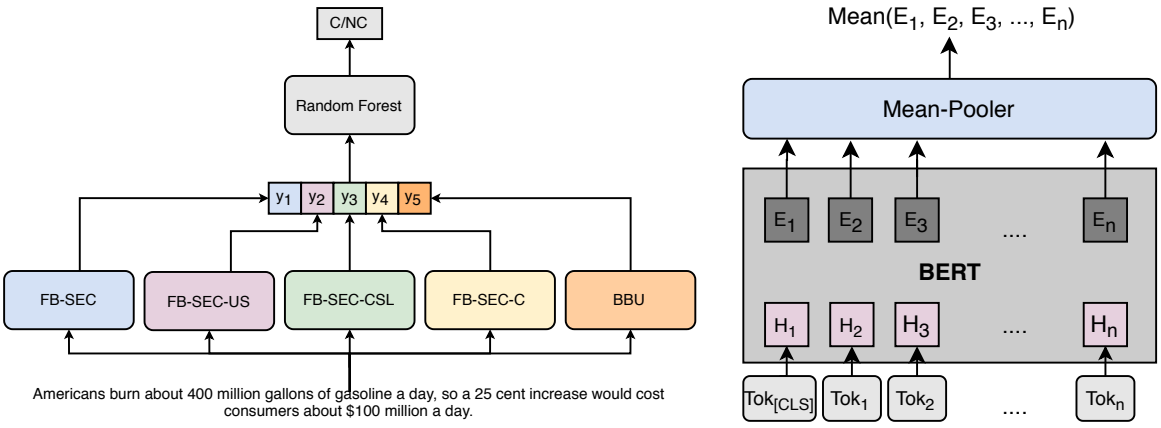

Figure 1: **Left**: Ensemble for the final system consisting of FinBert-SEC trained on task data (`FB-SEC`), FinBert-SEC with under-sampling on task data (`FB-SEC-US`), FinBert-SEC with Cost-Sensitive Learning on task data (`FB-SEC-CSL`), FinBert-SEC trained on task data augmented with Causaly without CSL or US (`FB-SEC-C`), and Bert-Base-Uncased trained on task data (`BBU`). *C/NC* refers to *Causal/Non-Causal* and y_n are the predictions by the individual models. **Right**: Demonstration of the Mean-Pooled embeddings calculated using the mean of the last-layer hidden states of the tokens

models to capture their strengths by using the predictions of the individual models as the features for training a second-order Random Forest classifier with 5 estimators and a maximum depth of 100. Our final system is shown in figure 1 (left).

3 Experiments and Results

We split the annotated corpus as 80% for training, 10% for validation, and 10% for testing. The final system is retrained on all the data combined for the number of training epochs obtained from the validation set using Early Stopping (Caruana et al., 2001) with patience of 3 epochs on the F1 score. We use PyTorch (Paszke et al., 2019) models trained on Nvidia GTX1050 GPU. We aim to answer the following Research Questions (RQ) using Precision (P), Recall (R), Weighted F1-score (F1), Confusion Matrix (TP, TN, FP, FN) on our own test set (columns without *-E) and official evaluation set (columns with *-E).

3.1 RQ1: Does further pre-training BERT on financial text affect the Causality Classification performance?

We hypothesize that the BERT models benefit from further pre-training on the financial text corpus before fine-tuning for causality classification. Therefore, we compare the BERT models mentioned in section 2.2, all of which are trained with identical hyperparameters: (i) Maximum Sequence Length of 200, (ii) Batch Size of 8 with Gradient Accumulation of 2 steps, (iii) Adam Optimizer with $4e^{-5}$ Learning Rate that decreases linearly after linearly increasing during a warm-up period of 335 steps. Table 2 shows that the `FB-SEC` outperforms the other BERT models w.r.t. the weighted F1-score on our own test set and is therefore used as the baseline for comparison in next research questions.

3.2 RQ2: What are the effects of training strategies such as undersampling, data augmentation, and cost-sensitive learning on the classification performance?

We try several strategies to tackle the imbalanced target label distribution.

Undersampling (US): We reduce the number of *non-causal* instances in the training set so that the proportion of the two classes is equal (50-50). Table 3 shows an improvement in the detection of the *causal* class through TP and FN.

Data Augmentation (DA): We select randomly the training instances from the augmenting corpora mentioned in section 2.1 and concatenate with the training set maintaining the target distribution. We see

Model	P	R	F1	TP	TN	FP	FN	P-E	R-E	F1-E
BERT-base-uncased (BBU)	94.42	95.01	94.34	68	**2028**	**20**	90	95.28	95.64	95.22
FinBert-FinCom (FB-FC)	93.39	94.28	93.40	56	2024	24	102	94.91	95.34	94.87
Finbert-TRC2 (FB-TRC2)	94.24	94.87	94.25	69	2024	24	89	**96.46**	**96.62**	**96.36**
Finbert-SEC (FB-SEC)	**95.19**	**95.55**	**95.29**	**91**	2017	31	**67**	96.19	96.42	96.23

Table 2: Comparison of generic BERT model with the Financial BERTs

Strategy	Configuration	P	R	F1	TP	TN	FP	FN	P-E	R-E	F1-E
Baseline	FB-SEC	95.19	95.55	95.29	91	2017	31	67	96.19	96.42	**96.23**
US	**FB-SEC 50-50 (FB-SEC-US)**	95.30	91.93	93.03	**140**	1888	160	**18**	95.11	90.90	92.23
DA	Training+SemEval (FB-SEC-SE)	94.79	95.01	94.89	95	2001	47	63	94.17	94.05	94.11
	Training+ADE (FB-SEC-ADE)	**95.34**	95.37	95.36	**106**	1998	50	**52**	94.79	94.81	94.80
	Training+Causaly (FB-SEC-C)	95.33	**95.64**	**95.43**	95	**2015**	**33**	63	96.18	96.26	96.22
CSL	**FB-SEC scaled (FB-SEC-CSL)**	95.62	95.92	**95.70**	**97**	2019	29	**61**	**96.22**	**96.43**	96.19

Table 3: Effect of training strategies on the performance

from Table 3 that augmenting the training set with *Causaly* (FB-SEC-C) outperforms other augmentations w.r.t. the weighted F1-score on our own test set.

Cost-Sensitive Learning (CSL): We scale the loss resulting from incorrect classification of a *Causal* instance during the training of the model using scaling factor $\alpha = \frac{1}{C} * \frac{N(D)}{N(D_{causal})}$ where C is the number of classes, $N(D)$ is the number of samples in data D and $N(D_{causal})$ is the number of samples where label is *causal*. Table 3 shows that the CSL leads to more True Positives than the baseline as expected.

3.3 RQ3: Which models should be combined for an ensemble and how?

Table 4 shows that ensembling using a second-order Random Forest (RF) classifier performs marginally better than majority-voting, and that the overall performance of the ensembled system is higher than the individual models. Our final system is an ensemble of FB-SEC, FB-SEC-CSL, FB-SEC-US, FB-SEC-C, and BBU.

Configuration	Ensemble	P	R	F1	P-E	R-E	F1-E
Final system	Majority Voting	95.89	96.14	95.96	96.90	96.99	96.93
Final system	**Second-Order RF**	**96.10**	**96.19**	**96.14**	**96.95**	**97.03**	**96.98**
Final system w/o BBU	Majority Voting	95.57	95.92	95.54	96.60	96.52	96.56
Final system w/o BBU	Second-Order RF	95.82	96.05	95.89	96.51	96.68	96.55

Table 4: Comparison of various ensemble configurations

4 Conclusion

We proposed a system to detect causality in the financial texts using an ensemble of domain-specific and generic BERT classifiers, trained using various datasets and training strategies to tackle the imbalanced class distribution. Among many other experiments that did not improve the performance and are not mentioned in this paper are the Convolutional Neural Networks (CNNs), RoBERTa language model (Liu et al., 2019), and non-neural classifiers trained on hand-crafted linguistic features. We would like to continue the improvement by jointly training the Causality Classifier (sub-task-1) with the Causal Span Extraction (sub-task-2).

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models, 08.

Rich Caruana, Steve Lawrence, and C. Lee Giles. 2001. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 402–408. MIT Press.

Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic extraction of causal relations from text using linguistically informed deep neural networks. pages 306–316, 01.

Vinicio Desola, Kevin Hanna, and Pri Nonis. 2019. Finbert: pre-trained model on sec filings for financial natural language tasks, 08.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12*, MultiSumQA '03, page 76–83, USA. Association for Computational Linguistics.

Harsha Gurulingappa, Abdul Mateen, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, http://dx.doi.org/10.1016/j.jbi.2012.04.008, 04.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

C. Khoo, Jaklin Kornfilt, R. Oddy, and Sung-Hyon Myaeng. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, 13:177–186, 12.

Sanghee Kim, Rob Bracewell, and Ken Wallace. 2007. A framework for automatic causality extraction using semantic similarity. 01.

Manolis Kyriakakis, Ion Androutsopoulos, Artur Saudabayev, and Joan Ametllé. 2019. Transfer learning for causal sentence detection. pages 292–297, 01.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications.