NITK NLP at FinCausal-2020 Task 1 Using BERT and Linear models.

Hariharan R L
Dept of Information Technology
National Institute of Technology
Karnataka
hariharanrl.22@gmail.com

Anand Kumar M
Dept of Information Technology
National Institute of Technology
Karnataka
m_anandkumar@nitk.edu.in

Abstract

FinCausal-2020 is the shared task which focuses on the causality detection of factual data for financial analysis. The financial data facts don't provide much explanation on the variability of these data. This paper aims to propose an efficient method to classify the data into one which is having any financial cause or not. Many models were used to classify the data, out of which SVM model gave an F-Score of 0.9435, BERT with specific fine-tuning achieved best results with F-Score of 0.9677.

1 Introduction

The important aspect as far as the financial news is the variability and the impact which it causes. In the information retrieval process, causality is an essential and well-known topic. Several NLP methods can be used to find the relationship between financial data and its effect. The main focus of this work is to come out with a better solution for the FinCausal-2020 shared task (Mariko et al., 2020). This shared task mainly focuses on determining causality associated with the financial object's transformation in quantified facts.

We have applied classification for the financial data using Linear Model and Deep Learning BERT (Devlin et al., 2018) model. In the case of Linear model, an SVM classifier is used, which is further fine tuned to produce better result. The fine-tuned BERT base uncased version was used as a deep learning model.

This paper is presented as follows; the details about the data being used are explained in section 2, system description is being explained in section 3, results and discussion in section 4, which follows the conclusions in the last section 5.

2 Dataset Description

The task organisers provided the data for the shared task as CSV files, namely trial, practice, and evaluation. These data where extracted from a corpus of 2019 financial news provided by Quam. The original data being HTML pages corresponding to the daily financial news feed is extracted. These raw set is being arranged with the column as Index, Text, and Category. Initially, the trial and practice dataset were released to build and train the model, which consists of data as shown in the table 1. The trial data had 8580 sentences with labels indicating whether there is any causality(1) or not(0), similarly 13478 sentences for practice data. The evaluation data had 7386 sentences without any labels and needed to be evaluated and appended with the prediction labels.

Example Sentences from the Dataset:

- Virtually free comprehensive medical care would lead to big increases in the demand for services:0
- Transat loss more than doubles as it works to complete Air Canada deal:1

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 60–63
Barcelona, Spain (Online), December 12, 2020.

Data	# of Sentences	Category	
		0(No causality)	1(Causality)
Trial	8580	8011	569
Practice	13478	12468	1010

Table 1: FinCausal 2020 Dataset details

3 System Description

We have developed both linear as well as deep learning models. The SVM classifier was used as a linear model and, BERT was used as a deep learning model. The hardware used for the experiments were Colaboratory by Google with GPU ranging from 10GB to 16GB(Tesla K80/Tesla P100). The description of each system is explained one by one below.

The Linear model SVM classifier was given as a baseline by the task organiser. We have tried with NBSVM model (Wang and Manning, 2012), To apply SVM model for the textual data, some basic preprocessing like removing URL, HTML tags (Richardson, 2007), special symbols, and accented characters were done. Further, these texts were converted to lowercase and TF-IDF vectorizer (Salton and McGill, 1986) was applied. The experiment was conducted in two phases, the first one being training the model using the trial dataset and testing with the practice dataset. The second one was training the model using practice dataset and testing with the trial dataset. The same steps were followed for both the phases. The prediction for the evaluation data provided by the organiser is done by building a model which was trained on both practice and train data.

The first phase of the experiment was done by splitting (Pedregosa et al., 2011) the trial data into 85% and 15% for train and validation, respectively. The TF-IDF was experimented with different n-grams and fixed a range of (1-5). The minimum and the maximum number of occurrences (min_df and max_df) of words to be considered to make a vocabulary were also altered. After grid search, the maximum and minimum occurrences were fixed at 90% (maximum number of sentences) and 2(least number of sentences), which gave better scores for the metrics as in table 2. The same process was repeated for the practice data, which was used to predict the trial data.

The practice and trial data were combined to predict the evaluation data whose scores are also given in table 2.

Task	Trained using	F1	Recall	Precision
Predict Practice Data	Trial	0.926486	0.940867	0.934973
Predict Trial Data	Practice	0.939693	0.943590	0.937448
Predict Evaluation Data	Both	0.943532	0.948687	0.943193

Table 2: SVM Model Results for Trial, Practice and Evaluation Data

The BERT model (Devlin et al., 2018) was used as a deep learning model for the classification task which is based on transformers (Wolf et al., 2019). Here we have used the BERT-base uncased pretrained model and trained FinCausal data on top of it. As the BERT model don't require any preprocessing, it wasn't done. Initially, during the evaluation phase, the BERT model was directly applied with the practice and trial data without any fine-tuning which gave us a result lower than that of the SVM model as shown in table 3.

Task	Trained using	F1	Recall	Precision
Predict Practice Data	Trial	0.868717	0.876911	0.860889
Predict Trial Data	Practice	0.865269	0.870159	0.860506
Predict Evaluation Data	Both	0.851731	0.846872	0.856710

Table 3: BERT Model Results for Trial, Practice and Evaluation Data

The results shown above are done before the evaluation deadline. Post evaluation, the model was

fine-tuned with changing important parameters as given below.

- Batch Size: kept as 6 for both training and validation
- Epochs: varied with early stopping keeping lesser validation loss
- Neither attention nor segments were maintained
- Learning rate: Kept at $2e^{-5}$.
- The LR parameter was tried with one fit one cycle and auto-fit learning rate using learning policies (Smith, 2017).

Task	Trained using	F1	Recall	Precision
Predict Evaluation Data	Both Practice and Trial	0.967770	0.967100	0.968712

Table 4: BERT Model Result Post Evaluation Deadline

The cyclic learning rate policy was used as mentioned in (Smith, 2017). This method was adopted for evaluation data, which helped to tune the learning rate and showed improved result than the other two models as shown in table 4.

4 Results and Discussion

Here we will explain the results obtained for the two phases of experiments conducted and the result that the model has given for the evaluation data. As shown in tables 2 and 3, both the models with modified parameters gave better results. These models were used to predict the evaluation data. The leaderboard after the evaluation deadline along with our updated score is as given in table 5. Our result was at 9^{th} position, which was the result obtained from the linear SVM model, as mentioned earlier. On further exploring the BERT model, we could get better results, which shows our proposed fine-tuned BERT model score near the 6^{th} position in the final leaderboard.

Team	F1	Recall	Precision
NITK NLP[1]	0.967770 (6)	0.967100 (6)	0.968712 (6)
NITK NLP	0.943532 (9)	0.948687 (9)	0.943193 (9)

Table 5: Leaderboard Positions

Hence, the results show that BERT model performed if the hyperparameters were well-tuned as it had a large corpus of pretrained data. After the evaluation, it was evident that the model could perform better as the accuracy difference with the first place and the accuracy improvement over the earlier BERT.

5 Conclusions

The FinCausal-2020 shared task was mainly aimed to come up with a model that could analyze the text and say whether it belongs to a particular financial causality or not. The main challenge was the imbalanced dataset and from which we need to develop a model that could produce an accurate result. As per the experiments being conducted and observing the results, the fine-tuned BERT model could perform well for the blind dataset using the k nown data than BERT and the liner SVM model. If we could come up with more balanced data with sampling methods, BERT would outperform most of the existing models.

[1]Post Evaluation

References

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*.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Leonard Richardson. 2007. Beautiful soup documentation. *April*.

Gerard Salton and Michael J McGill. 1986. Introduction to modern information retrieval.

Leslie N Smith. 2017. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE.

Sida Wang and Christopher D Manning. 2012. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. Technical report.

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.