BERT for Financial Domain FinBERT: Replication & Extension

Khevna Parikh
kp2936@nyu.edu

Lakshana Kolur
lk2719@nyu.edu

Tanya Naheta
tsn6109@nyu.edu

Abstract

Information about financial data is highly valuable as it gives insight into the company's revenue, growth and risks which is beneficial to companies, investors and the industry as a whole with the power to move markets. This led to the development of FinBERT, a state-of-the-art language model based on Google's BERT that adapts specifically to the finance domain. In this paper, we aim to replicate the FinBERT model to analyse the sentiment behind financial texts and understand where it may have room for improvements. We, then, fine-tune the model on different datasets to improve sentiment analysis in the areas that the model failed to perform well. We also extend the model to identify ESG goals of publicly-traded US companies. Furthermore, examine how FinBERT performs as compared to the BERT base model and if training on financial data makes a difference.

1 Introduction

As new natural language processing (NLP) technologies and tools have developed in past decade, analysis of financial texts has become of an essence for organizations to remain prominent in the market. Determining the underlying state or tone of the market early on could provide insights into events, such as the rise or fall of a particular security, allowing an investor to predict profitability from such assets. Hence, the key research interest of this thesis is the sentiment analysis of financial texts. On the contrary to traditional sentiment classification of labeling texts as positive or negative, financial sentiment analysis focuses instead on predicting how the markets react from the information provided in a given text, be it from an official company announcements or analyst reports.

Most traditional classification methods require large labeled financial texts, costing financial entities time and money. Furthermore, these models are not well-suited for the financial domain as the semantic meaning of industry jargons would differ in context. For example, generally the term risk would mean harm or danger but in finance, it refers to the monetary consequence of a stock option or a security. To combat this issue of algorithms not performing well specific to domain, Dogu Araci fine tunes a state-of-the-art language model on large corpus of unlabeled financial texts including corporate filings, analyst reports, and earnings conference call transcripts, introducing FinBERT, a nerual language model trained purely on financial corpus.

The Bidirectional Encoder Representations from Transformers (BERT) algorithm is trained on large amounts of general text, learning the semantic and syntactic relations between words. This context-specific vector space is outlined clearly in an example by the author - "bank" is learned as more closely related to "lending" rather than "fish" or "river" (Araci, 2019). Put simply, FinBERT is a variation of the BERT based language model; hence, it is important to understand what the BERT model entails. Thus, the main contributions of this thesis would be: (1) to introduce the architecture of the BERT model, (2) to replicate Araci's work in solely training a BERT model on financial corpus in hopes to achieve similar accuracy for evaluating financial sentiment analysis task, and (3) conduct experiments to further fine-tune on additional datasets to improve sentiment classification on the mislabeled sequences (4) extend the model to identify ESG goals in financial company documents.

1.1 Model Architecture of BERT

Neural networks can only understand numerical digits so the initial step includes assigning numerical values to all words in each sentence. Consider the following sentence, "The price of APPL will drop tonight". The BERT tokenizer will firstly transform a sentence into a sequence of words, or tokens. Additionally, it will append two special tokens (1) [CLS], the first token of every sequence, representing a classification token and (2) [SEP],the last token of each sentence, indicating with token belongs to which sequence. The maximum size of tokens possible is 512. If the sequence is lesser in size, we pad using the [PAD] token and if the sequence is longer, we truncate accordingly. Along with the token embeddings, BERT encompasses positional embedding (position of the tokens) and segment embeddings (used for sentence pairs) as well. This embedding become the input of the BERT model.

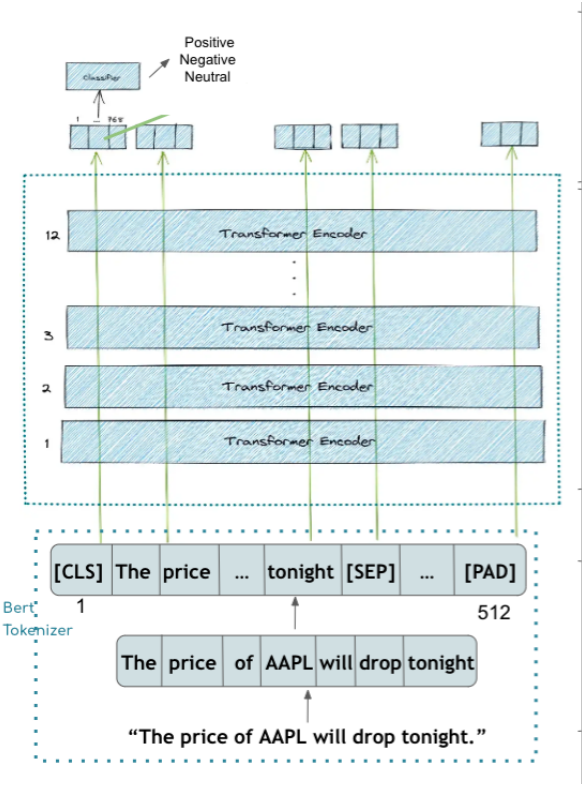


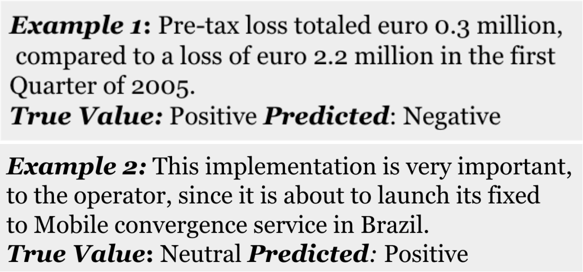
Figure 1: BERT Architecture

Remarkably, the model learns information from a sequence of tokens in both directions, from left to right and right to left, hence Bidirectional in the name. The BERT base model consists of 12 layers of Transformer encoders stacked together, an attention mechanism that learns contextual relations between words (or sub-words) in a text. It decides at each step of an input sequence which other parts of the sequence are important. Lastly, it trains using two strategies, (1) Replaces certain tokens with [MASK] to allow for the prediction of the original value of the masked words, and (2) predict if the second sentence is subsequent sentence in the original document and if they are related in context. BERT can successfully be applied to tasks such as question answering or next sentence prediction. In the next section, we talk about the most-used case, sentiment classification.

2 FinBERT: Sentiment Analysis

2.1 Data Collection

To replicate Araci's results, we collected the labeled Financial PhraseBank dataset[1]. Financial PhraseBank consists of 4,846 randomly selected English financial news statements annotated by 16 individuals with an extensive background of financial markets. The annotators were asked to give labels according to how they think the information in the sentence might affect the mentioned company stock price. The dataset also includes information regarding the agreement levels on sentences among annotator; sentences with agreement of greater than 50% were used to train the BERT model.


***Example 1*:** Pre-tax loss totaled euro 0.3 million, compared to a loss of euro 2.2 million in the first Quarter of 2005.
True Value: Positive ***Predicted***: Negative

Example 2: This implementation is very important, to the operator, since it is about to launch its fixed to Mobile convergence service in Brazil.
True Value: Neutral ***Predicted****:* Positive

Figure 2: When does BERT fail?

2.2 Training

We fine-tuned a pre-trained BERT model on the above dataset to classify a specific financial text as either neutral, positive, or negative. By means of the BertTokenizer class, we transformed our sentences into a sequence of tokens that the model was easily able to interpret. After pre-processing, we trained the model using the same configurations employed by Araci, as shown in figure 2. This included imposed Adam as the optimizer with a learning rate of 2e-5 and utilizing cross entropy as our loss function.

After 4 training epochs, we were able to achieve a training accuracy of 0.76 and a validation accuracy of 0.74. Even with using the same model configurations, these results were not in par with the 86% accuracy retrieved by Araci. With an additional epoch run, we were able to reach a training accuracy of 0.82. While keeping all hyper parameters the same, we ran the model on our hold-out sample and received an accuracy of 0.70 with precision of .66. In other words, 66% of the time, our model correctly classifies the positive sentiments. We find that even with a smaller training set and training for a short amount, FinBERT outperforms state-of-the-art machine learning methods.

These results can be quite prominent, however there are a few cases when FinBERT fails to properly classify labels as seen in the examples above.

[1] https://huggingface.co/datasets/financial_phrasebank

It has difficulty distinguishing words of directions, such as increase or decrease, especially in the context of mathematical figures. It further has a difficulty understanding sentences when there is a lack of company information available as we see in example 2. This motivates us to further investigate and reduce miscategorization of positive, neutral, negative statements due to inflated outlooks by financial institutions?

3 Use semi-supervised learning to improve positive-neutral classifications

The goal of this extension was to improve the performance of the FinBERT sentiment classification between positive and neutral statements by introducing additional labeled datasets, as this was one of the main weaknesses outlined by Araci. However, one of the biggest challenges with developments in this area is the lack of these labelled datasets in the finance domain. We were able to bypass this problem by performing semi-supervised learning on an unlabeled dataset.

3.1 Data

We used the News Aggregator Data Set by University of California, Irvine that contains headlines for 152,746 news stories in the Business cateogry collected by a web aggregator between March 10th, 2014 and August 10th, 2014 (Dua and Graff, 2017).

3.2 Experiment setup

To focus specifically on the types of positive-neutral statements that the model labelled incorrectly, we retrieved and analyzed these sentences and conducted a similarity search to find similar sentences from the news headlines dataset. Next, we added the sentiment labels for the headlines that had similarity scores greater than 0.80 using the original FinBERT model (551 statements), and concatenated this to the PhraseBank labeled set to reinitialize and train our model. We use Sentence-Transformers, a PyTorch framework with several pre-trained models specifically trained for semantic search (Reimers and Gurevych, 2019).

3.3 Results

Extending the FinBERT model with additional data resulted in higher accuracy, precision, and f-1 scores as can been seen in Table 1. Specifically, we reduced the number of incorrectly labeled positive-neutral statements by 11.8% (from 178 incorrect statements to 157). However, this extension will have similar biases to the original FinBERT model as that is what was used to label the new dataset. Using fully labeled datasets would allow us to remove these biases and improve our model.

Model	Accuracy	Precision	F1-score
FinBERT	0.71	0.66	0.61
FinBERT + News	0.81	0.75	0.62

Table 1: Evaluation metrics for FinBERT vs. News Data Extension

4 FinBERT: Identify ESG Goals

The second extension of the FinBERT model was to identify if publicly-traded US companies spoke about their ESG (environmental, social and governance) goals, specifically, about their greenhouse gas emissions reduction targets. The motivation for this extension lies in understanding that ESG goals have a direct impact on the company performance, i.e. studies show that companies with higher ESG ratings are associated with higher stock prices, lower market volatility and higher sustainability. Hence, identifying such sentences (referred to as ’target sentences’) can be of great importance for a company.

4.1 Data collection

Since the focus is publicly traded companies, the data sources were publicly available company SEC filings. This included 10-Ks, 10-Qs, 8-Ks and proxy statements. The main issue with these documents was that although the data was abundant, the actual target sentences were of a very small proportion. It was a highly imbalanced dataset with 466,000 sentences but only 210 target sentences. For the model to give plausible results, increasing the number of targets was necessary. This was carried out by selectively replicating the target sentences. What we found to be effective was that around 1000 target sentences in the training set with a total dataset size of 466,000 was sufficient to effectively train the model.

4.2 Experiment setup

For this experiment, the trained FinBERT model was the baseline since the model already had exposure to financial documents. This was further trained on SEC filings to specifically comprehend target vs. non-target sentences. The overall setup and architecture of the model was fairly similar to the FinBERT base model with the difference on what they were both predicting - the FinBERT model was predicting sentiment whereas the extended model was predicting if a sentence contained ESG information or not.

4.3 Results

Since the dataset was enormous, training on all 466,000 sentences took roughly 3 hours for 1 epoch leading to different training experiments that will be detailed in the table below. With imbalanced classes, it's easy to get a high accuracy without actually making useful predictions. Hence, for this experiment we focused more on the precision and recall metrics for the under-represented target class.

To see how well the FinBERT model performs and if the training on an abundant amount of financial text is beneficial, the same experiments were performed on the BERT base model. The metrics can be seen in the table below.

Model	Precision	Recall	F1-score
BERT-base			
Sample	0.48	0.42	0.45
50% dataset	0.69	0.32	0.44
100% dataset	0.27	1.00	0.43
FinBERT			
Sample	0.74	0.43	0.55
50% dataset	0.80	0.74	0.77
100% dataset	0.80	0.81	0.79

Precision refers to the exactness of the model, or how accurate are the positive predictions. We notice that FinBERT has higher precision as compared to all the experiments conduced on BERT-base. This shows that our model is able to accurately identify the correct target sentences despite the replication. This can be attributed to the fact that since the model has exposure to financial jargons, it is able to correctly distinguish between true positives and false positives. Recall, on the other hand is a measure of the completeness of the model. It refers to the coverage of the actual target samples. Although the BERT base model has a 100% recall, it is at the cost of a low precision. This is indicative that it managed to identify all the target sentences, but at the cost of too many false positives. FinBERT, on the other hand, does not have such a high recall, but maintains the precision and aims to achieve a high recall, high precision outcome resulting in a high F1-score.

5 Conclusion

Attributing to its extensive financial relevance, we understood the working of BERT and replicated Araci's FinBERT. Despite keeping the same hyper-parameters, the results were not at par with the paper. However, we found that the model performed well with smaller training sets and for a shorter time. We analysed where the model was failing & fine-tuned the model on additional labeled datasets to improve sentiment analysis. Additionally, we extended the model to identify companies' ESG goals and observed that FinBERT outperformed the BERT model due to its adaptation to the financial domain.

6 Future work

In order to tackle model failures due to numerals, an extension to classify whether numerals in a statement are in-claim or out-of-claim would be very beneficial. However, as outlined earlier, the biggest challenge we foresee is the lack of labeled datasets by finance domain researchers that can be used to train a model. Furthermore, by using a NER task to identify, categorize, and encode financial entity information, we can extract important prior knowledge about the entity which can be learnt jointly with original statements.

7 Author Contribution Statement

Khevna Parikh: Equally contributed to project proposal, poster and final report, 100% contribution for progress update report and replication of the FinBERT sentiment analysis. Lakshana Kolur:Equally contributed to project proposal, poster and final report, ran experiments for the progress update, 100% contribution for Extension 2. Tanya Naheta: Equally contributed to project proposal, poster and final report, 100% contribution for Extension 1.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063.

Dheeru Dua and Casey Graff. 2017. UCI machine learning repository.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.