BERT for Financial Domain: FinBERT
Sentiment Analysis & Reduce Miscategorization

NYU Center for Data Science

Khevna Parikh, Lakshana Kolur, Tanya Naheta

Abstract

Analyzing financial text sentiment is valuable as it can engage the views and opinions of managers, information intermediaries and investors. FinBERT is domain specific state-of-the-art BERT model, fine-tuned on corporate filings, analyst reports, and earnings conference call transcripts.

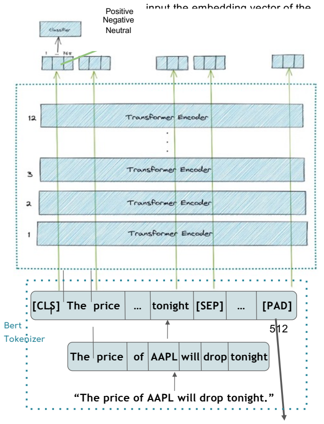
Goal: Replicate and fine-tune FinBERT to improve sentiment analysis of financial statements & extend FinBERT

Methodology:
- ❑ Use a pre-trained BERT model to categorize financial-based texts into one of the following labels: positive, negative or neutral
- ❑ Fine-tune FinBERT on additional data to improve sentiment classification
- ❑ Extend FinBERT to identify sentences in US companies regarding their ESG (environmental, social, governance) goals.

What is BERT?

Bidirectional **E**ncoder **R**epresentations from **T**ransformers
- ❑ Consists of several Transformers encoders stacked together, an attention mechanism that learns contextual relations between words (or sub-words) in a text
- ❑ Uses two training strategies:
  1) Replace certain tokens with [MASK] to allow BERT to predict the original value of the masked words`
  2) Predict if the second sentence is the subsequent sentence in the original document

For the classification task, further input the embedding vector of the



Maximum size of tokens possible is 512. If the sequence is less, we pad using the [PAD] token and if the sequence is longer, we truncate.

Output: Embedding vector of size 768 for each token.
- ❑ A probability of each word in the vocabulary using softmax.

BERT Base:
- ❑ Along with token embeddings, BERT uses positional embedding (position of the tokens) and segment embeddings (used for sentence pairs) as well.
- ❑ Consists of 12 layers of Transformer encoder, 12 attention heads, 768 hidden size, and 110 M parameters
- ❑ Attention mechanism decides at each step of an input sequence which other parts of the sequence are important
- ❑ Learns information from a sequence of tokens in both direction, from left to right and right to left

Input Tokens: Transform a sentence to a sequence of words, or tokens and add two special tokens
- ❑ [CLS]: first token of every sequence, representing a classification token
- ❑ [SEP]: last token of the sentence, indicating with token belongs to which sequence

This become the input of the BERT model.

Baseline Replication: Financial Sentiment Analysis

- ❑ Financial Phrasebank Dataset: 4,846 sentences from English financial news, annotated by 16 individuals with background knowledge of financial markets; includes instances with 50% annotator agreement

> Sentence: Lifetree was founded in 2000, and its revenue have risen on an average by 40% with margins in late 30s.
> Label: Positive

- ❑ Configurations used:
  (same as Dogu Araci's paper[1])

> Tokenizer: bert-base-cased
> Bert Model: bert-base-cased
> Num Train Epochs = 4
> Max Sequence Length = 64
> Train Batch size = 32
> Learning rate = 2e-5
> Output mode = classification

- ❑ Results: Unlike Araci, it took us 5 epochs to achieve similar training accuracy. All other hyperparameters remind the same.

> Epoch: 5
> Average training acc: 0.82
> Average validation accuracy: 0.76

> Test results:
> Accuracy 0.7061855670103093
> Precision 0.6594896331738437
> F1 Score 0.6140012738285491

Where does FinBERT fail?

1. Mathematical figures used in statements in the absence of directional words:
   ***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

Difficulty distinguishing between numbers, and words such as increased or decreased.

1. Lack of information about the company:
   Example 1: 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

Can we reduce miscategorization of positive, neutral, negative statements due to inflated outlooks by financial institutions?

Experiments

Extension 1: Fine-tuning FinBERT to improve sentiment classification using semi-supervised learning

- ❑ Data: 152,746 unlabeled news headlines from Business category over a 4 month period[2].
- ❑ Semantic similarity search with incorrectly labeled sentences by FinBERT
- ❑ Label the data using FinBERT model with similarity score higher than 0.8

Phrasebank sentence:

> "Kemira will supply the additional volumes of iron coagulants from the company's plant at Helsingborg , Sweden."

Similar sentences:

> "WeSchool matches 11 Swedish renewable energy companies with Indian partners."
> "Laval firm Valeant increases offer for Allergan, makers of Botox."

Extension 2: Extending FinBERT for identifying if companies have ESG goals, specifically, greenhouse gas emissions reduction targets.

- ❑ Dataset: SEC filings including 10-Ks, 10-Qs, 8-Ks of publicly traded US companies. Imbalanced dataset containing 466,000 sentences with 200 target sentences. For better performance, increased target sentences from 200 to 1000 by selective replication.
- ❑ FinBERT model modified for binary classification task: identify target & non-target sentences.
- ❑ Training experiments: (a) 50% of data (b) 100% of data (c) comparison of FinBERT with BERT Base.

Sentence	Target
We are on a path to powering our operations with 100% renewable energy by 2025—five years ahead of our original target of 2030.	1

Conclusions and Future Steps

Replication: In our attempt to replicate Araci's work, we were able to achieve a noteworthy accuracy in a very limited time. We noticed similar discrepancies is categorization of certain sentences, which further motivated us to investigate and reduce the mislabeled sequences.

Extension 1: Adding additional headlines from the news aggregator dataset increases precision. This is key as we have reduced false positives, addressing our research question.

```
Accuracy 0.8057395143487859
Precision 0.7506945550423811
F1 Score 0.624587798500842
```

Extension 2: FinBERT performs better than BERT on SEC filings, even for identifying ESG goals.

	precision	recall	f1-score
0	1.00	1.00	1.00
1	0.20	0.67	0.31
accuracy			1.00
macro avg	0.60	0.83	0.66
weighted avg	1.00	1.00	1.00

Low precision, high recall (actual target sentences being marked correctly)

Overall conclusions:
- ❑ FinBERT is a promising state-of-the-art sentiment analysis tool in the financial domain. It could allow companies predict a rise or fall in that company's stock.

Further extensions to FinBERT:
- ❑ Using labeled datasets to classify whether numerals in a statement are in-claim or out-of-claim.
- ❑ Using NER to identify, categorize, and encode entities in the financial domain. This could increase sentiment accuracy as extract important prior knowledge about the entity and is learnt jointly with the original texts[3].

References

[1] Araci, Dogu. "Finbert: Financial Sentiment Analysis with Pre-Trained Language Models." *ArXiv.org*, 27 Aug. 2019, https://arxiv.org/abs/1908.10063.

[2] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.