Separating Financial Fact from Opinion using Bidirectional Encoder Representations from Transformers (BERT)

Chae Yeon Lee, Aaron Chow, Jackson Fu, Ye Joon Han

Advisor: Adam Badawi (Berkeley Law)

Abstract

In recent years, machine learning algorithms such as ELMo and BERT have become increasingly popular for analyzing large volumes of textual data. One promising application of these algorithms is in the field of finance, where news articles and social media posts can provide valuable insights for investors. In this study, we trained a finBERT model and achieved a testing accuracy of 0.93. Our analysis revealed that the model places the most weight on words such as "may," "believe," and "could" when classifying a sentence as an opinion, and objective words such as "products," "including," and "revenue" when classifying a sentence as a fact. We also found that there were significantly more opinions in the MD&A section than the risk factor section. Interestingly, we observed a reduction in opinions after 2015, which was the year when the U.S. Supreme Court clarified liability for opinions in registration statements. We observed correlation between closing price and composition of fact, opinion, and neutral statements for MD&A and Risk Factor sections.

Background

Previous studies have shown that language models, such as ChatGPT, can effectively analyze financial text to extract sentiment, predict stock prices, and classify news events. Additionally, recently researchers have deployed the artificial intelligence chatbot in market-relevant tasks, such as determining whether Federal Reserve Statements are Hawkish or not. However, there has been limited research on using these models to distinguish between factual and opinionated financial text. The motivation for this research is to address this gap in the literature by utilizing BERT to develop a model that can accurately differentiate between factual and opinionated information in public filings. The two primary sources of textual information about the performance of public firms are the EDGAR database of public filings maintained by the Securities and Exchange Commission (SEC) and the transcripts of earnings calls from public firms. By developing the BERT model to analyze these financial documents, we aim to develop a tool that can assist them in making more informed investment decisions by identifying sources of bias and potentially misleading information.

Objective and Research Question

The goal of this project is to develop a machine learning model that identifies and separates the factual content of SEC filings from opinions expressed in them. We also aim to analyze earnings calls and identify specific language patterns that are associated with positive or negative stock price movements using the model. Also, we aim to understand the change in the composition of the fact and opinion in filings for different countries over time in order to understand whether and how changes in the legal consequences for stating facts and opinions have changed the way firms speak about their performance.

Materials and Methods

Data Collection Data was extracted using the SEC API. Datasets with the words "believe" or "belief" were collected for fine-tuning. We initially began by classifying sentences into two categories – opinion or fact, but added an additional category "neutral" for ambiguous sentences. We also collected daily historical stock data for five companies of focus – Apple, Microsoft, Tesla, Amazon, and Google – from Yahoo Finance for the last ten years.

BERT Model Development We began with a pre-trained NLP model called FinBERT which is built by further training the BERT language model for the financial domain specifically. FinBert was developed by Prosus, a global consumer internet group, and it is mainly used to analyze sentiment of financial text. We further trained the FinBERT model for our specific task by training on a corpus of about ~650 manually labeled sentences from SEC filings from several hand-picked companies. We evaluated the model using Integrated gradients, which compute the attribution of each feature in the BERT model based on the gradient of the prediction with respect to the input. They are widely applied to deep learning methods for classification and regression tasks. Green, white, and red highlights represent opinion, neutral, fact, respectively, and the darker the color is the more emphasis that the model puts on the feature to make prediction.

Model Deployment We then created a web service to deploy the model locally using Flask. However, to make this model more accessible to the general user, we are under the development of the web service over Heroku.

Results

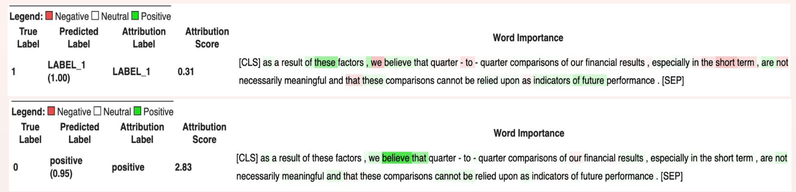

Figure 1-1. Integrated Gradients to Interpret BERT Model's Prediction prior to training our finBERT model with our training dataset.

Figure 1-2. Integrated Gradients to Interpret BERT Model's Prediction after training the finBERT model with our 900 training data points.

```
Loss:0.17
Accuracy:0.97

Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.98      0.97       337
           1       0.98      0.95      0.96       254
           2       0.88      0.98      0.93        45

    accuracy                           0.97       636
   macro avg       0.94      0.97      0.95       636
weighted avg       0.97      0.97      0.97       636
```

Figure 2. Training Metrics The model has been trained on a training dataset of size 636. 0, 1, 2 each represent opinion, fact, and neutral categories. F1-score is a harmonic mean of precision (fraction of true positives) and recall (fraction of true positives out of the total number of samples that belong to that class).

```
Loss:0.47
Accuracy:0.93

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.98      0.95        94
           1       0.97      0.90      0.93        70
           2       0.83      0.77      0.80        13

    accuracy                           0.93       177
   macro avg       0.91      0.88      0.89       177
weighted avg       0.93      0.93      0.93       177
```

Figure 3. Testing Metrics. The size of test dataset and validation dataset size are 177 and 71, respectively. 0, 1, 2 each represent opinion, fact, and neutral categories.

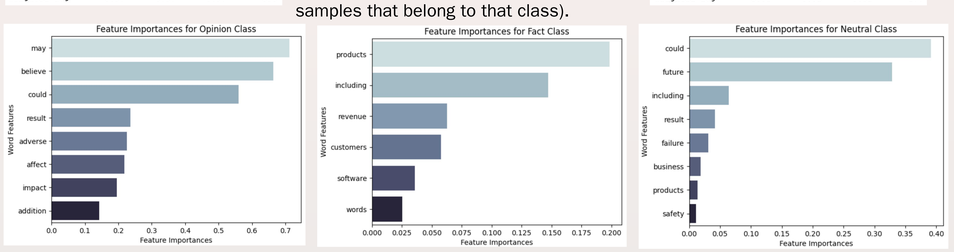

Figure 4. Feature Important for Each Opinion, Fact, and Neutral groups. These are feature importances generated from running integrated gradients on the BERT Model's prediction over the test set after removing stop words. Attribution scores of each word token are averaged over each predicted class.

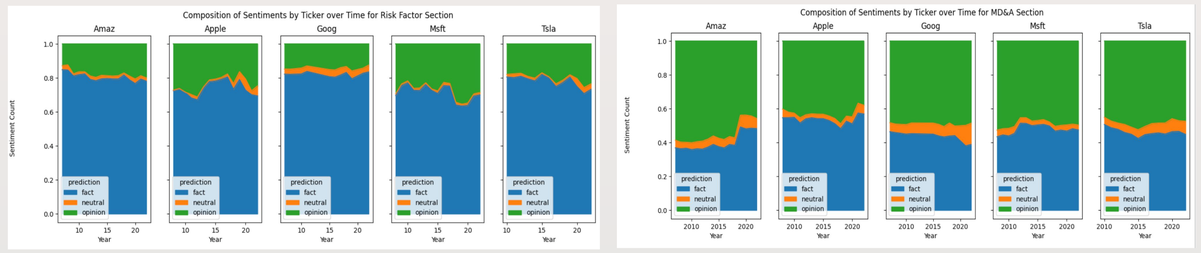

Figure 5. Composition of Sentiments by Ticker over Time for Risk Factor and MD&A section. Collected all available 10K public filings from year 2008 to 2023 for five companies – Amazon, Apple, Google, Microsoft, Tesla – and classified them into fact, opinion, and neutral with our trained language model. Upon visualizing, we selected Management's Discussion and Analysis of Financial Condition and Results of Operations section (MD&A) where the company's performance is analyzed, and Risk Factor section, where significant risks that the company faces are listed.

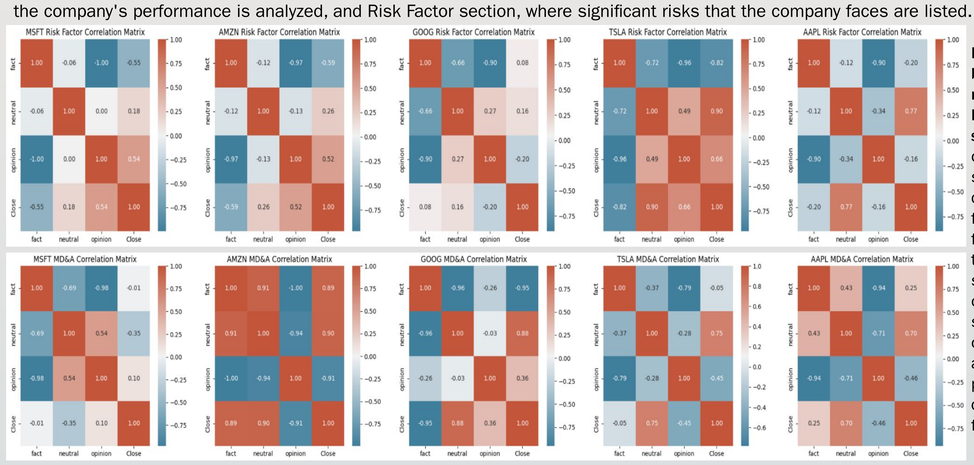

Figure 6. Correlation Matrix between the distribution of fact, neutral, and opinion and Stock Prices for Risk Factor and MD&A sections. Calculated the mean of open, high, close, low, close stock prices of the next five days after the release of public filings from year 2008 to 2023 for five companies. Measured the correlation between the stock prices and the proportion of fact, neutral, and opinion statements in the filing to observe if there is any association between the stock price movement and the composition of sentiments in filings.

Future Work

In the future, the model could be improved by training on more labeled data. Currently the model is only fine-tuned on about ~900 data points which is a still relatively small dataset. In addition there could be more work on hyperparameter tuning the model to ensure the most optimal hyperparameters are chosen to maximize performance. We also would like to develop a finBERT model with more prediction categories so that it incorporates sentiments with partiality of statements; i.e. the narrowed categories are fact-positive, fact-negative, neutral-positive, neutral-negative, opinion-positive, and opinion-negative. We hypothesize that sentiment has correlation with partiality of a statement, and thus studying them together allow us to perform more diverse research about how factuality of statements correlates with sentiments. Lastly, we hypothesize that reductions in opinions in public filings after 2015 for some companies may have some associations with 2015's Supreme Court case that clarified Liability for Statements of Opinion in Registration Statements. We plan to conduct statistical analysis examining potential causation or associations between them.

Interpretation

Fig 1 Integrated Gradients to Interpret BERT Model's Prediction. Prior to training the language model with our training datasets, the model puts most emphasis on the word "these" to predict that the sentence is opinionated. After training with our manually classified datasets, the model focuses on the word "believe" and correctly classifies the sentence as opinion. Furthermore, the attribution score increases from 0.31 to 2.83, which suggests that the model is more certain that the sentence is opinion. This figure shows that the training dataset has improved predictive accuracy of the model for opinionated sentences.

Fig. 2, 3 Training Metrics and Testing Metrics. The model's prediction accuracy on training dataset is 0.97, which indicates that the model give precise result for the data that it has seen before. When the model is tested on the testing dataset, which it has never encountered before, the accuracy is 0.93, which suggests that the model can correctly classify the sentences into fact and opinion for new datasets. Model has as the f1 score of 0.80 for neutral category, suggesting that it has relatively more difficult time predicting sentences in the neutral class, and has the highest f1-score for opinion category by a f1-score of 0.95. A notable difference in the f1-score for neutral category between train and test datasets, suggest that we need more training for neutral statements.

Fig 4 Feature Important for Each Opinion, Fact, and Neutral Statement groups. The three most important features the model thinks in classifying a sentence as opinion are 'may', 'could', and 'believe'. This result confirms that the model correctly puts emphasis on the words that indicates opinion when determining whether the sentence is opinionated or not. For fact class, the top three important features are 'products', 'including', and 'revenue'. It is interesting that the model puts most weight on the common word 'products' when classifying statements. Lastly, the neutral class puts the most emphasis on 'could', 'future', and 'including'. It is important to note that both fact and neutral class put emphasis on word 'including'. In general, we can see some overlapping features between these classes, which suggest that additional fine-tuning on more data may be needed to make clearer distinctions in features between classes.

Fig 5. Composition of Sentiments by Ticker over Time for Risk Factor and MD&A section. Over five different companies, MD&A section has significantly more opinions than risk factor sections, which aligns with our initial assumption that when a company reflects on and forecasts their performance in MD&A the public filings included many speculative statements. Furthermore, more than half of the MD&A statements are classified as opinions, suggesting that the MD&A section is heavily speculative. We can see that the fraction of opinion decreases after 2015 for Amazon, Tesla, and Apple. There are significantly less neutral statements compared to fact and opinion statements.

Fig 6. Correlation Matrix between the distribution of fact, neutral, and opinion and Stock Prices for Risk Factor and MD&A sections For MD&A section, three of five companies show that opinion and closing price have negative correlations, and that there is unanimously positive correlation between neutral statements and closing price. However, there is no strong correlation between fact and closing price because the correlation between them significantly varies by companies without any overarching pattern. For Risk factor section, four companies show that fact and neutral statements have a positive correlation with closing price. There is no obvious association between opinion and closing price. Therefore, two sections – MD&A and risk factor – identify different correlation pattern between closing price and partiality of statements.

Conclusion

In conclusion, our study demonstrates the effectiveness of machine learning algorithms in analyzing textual data related to financial markets. Our finBERT model achieved a high testing accuracy of 0.93 and was able to accurately classify statements as opinions or facts based on the presence of certain keywords. We found that the MD&A section contained significantly more opinions than the risk factor section, which may reflect the subjective nature of the information presented in the MD&A. It is also interesting to note that there was a reduction in opinions after 2015, which may be related to the U.S. Supreme Court's clarification of liability for opinions in registration statements. Overall, our findings highlight the potential of machine learning algorithms for extracting valuable insights from financial text data and provide important insights into the characteristics of opinion and fact statements in financial filings.

Acknowledgement We would like to extend our gratitude to Professor Adam Badawi for guiding us through this research and Data Discovery Program for providing us this learning opportunity.