Sentiment Analysis for Financial Earning Calls

Elena Relic-Lytle, Shachar Erez, Shane Kalepp

elena.relic.lytle@gmail.com, shachare@gmail.com, skalepp@gmail.com

Northwestern University

March 7, 2021

Abstract

Our research was conducted utilizing sentiment analysis to help quantify public earnings calls done by public companies. Outlined in this report presents our findings and research results by using the sentimental value for each earnings call. Our objective was to create a model that can detect positive or negative connotations and subsequently determine the effects of the share price (pre and post-earnings call). We utilized various natural language processing (NLP) techniques such as bag-of-words, Bidirectional Encoder Representations from Transformers (BERT), and Financial BERT (FinBERT). The primary goal is to understand if a correlation between the share price and sentiment during the call has any effect on the post-share price, based on our chosen model.

Introduction

The goal of our research is to develop a minimum viable product using natural language processing to analyze various company's quarterly earnings reports. The opportunity we had with this research was to analyze the overall sentiment of the earnings call and how it correlates with the company's stock price. Over the duration of our research, much of the value we derived by doing exploratory research and analysis. We utilized various NLP techniques and specific finance datasets to help train our model to have more domain knowledge that's widely used in the financial industry. In our methodology section, we elaborate further, but these datasets will become imperative in training our models.

Literature Review

Analyzing alternative data such as NLP of quarterly earnings calls is an important topic to study as investors aim to synthesize relevant disclosures that firms release to the market

(McKay Price et al. 2012). The research conducted by McKay Price et al. (2012) focused on the tone of the question-and-answer section of the conference calls and analyzed trading performance for 60 days following the call. They accomplished their study by employing a general word categorization dictionary as well as a custom, earnings-specific dictionary to determine the sentiment of the earnings call (McKay Price et al. 2012). Our methodology built a baseline model using a bag-of-words approach, which we compared against a BERT model and a FinBERT model.

The first model we built was the BERT model developed by Google. This is a method that offers a unique way of breaking down each of the sentences as tasks. This particular method looks to use two strategies when initiating pre-trained language, feature-based and fine-tuning. The feature-based approach looks to include the tasks as added features, while the fine-tuning approach tries to add task-specific parameters (Devlin et al. 2019).

The FinBERT method was developed by Dogu Araci which looked to improve on the already existing BERT framework but give it financial expertise. The basis of his method is a three-step approach; the use of a subset of the Reuters TRC2 data set to pre-train the model, Phrasebank (smaller dataset) that labeled the news headings based on their annotation, and lastly the vanilla BERT method which uses no pre-training (Araci 2019).

Another methodology to transfer learning is Universal Model Fine-tuning (ULMFiT), which enables a robust transfer learning framework for many NLP use cases. ULMFiT is universal enough that it can be used across a plethora of domains. The strength of this model is it can be relied on for long-term dependencies, hierarchical relationships, and sentiment tasks

(Howard & Ruder 2018). We didn't use this approach as part of our minimum viable product, but would look to draw on this method for future work.

Data

Our minimum viable product involved intensive exploratory analysis to find specific patterns or trends that we saw and wanted to explore. Data cleaning was a large effort for this research as we had tasks such as removing stop words and labeling some of the data. The data collection process utilized a web scraper from SeekingAlpha.com and exported the HTML into a text file. We separated text files for each company for the given quarter. Our research was based on Q4 2020 data as the timing of this writing most companies had their Q4 calls.

We used a public dataset called Financial Phrase Bank including two columns, financial news headlines, and sentiment. We utilized this dataset as a training dataset to help improve the accuracy of our model for when we were running our test sets.

Methods

The first method we looked at was the bag-of-words model using Python and the NLTK package. This method is a rather simple representation of classic NLP techniques. This approach takes into account the sentence as a vector and places a numerical value for each word. It uses those values and counts the number of occurrences that the word appears in the vector. One of the major drawbacks to this approach is we don't gain any information about the context of the sentence. Our approach to algorithm selection will be a fluid and iterative process, often metrics like accuracy, performance, and feasibility will come into the decision-making process.

We decided to focus on employing more novel techniques such as FinBERT that was developed off of the language model BERT. This method looks to provide improved labeling

techniques specifically for the financial domain. During our research and tests, we found this to be one of the more challenging aspects of conducting NLP experiments. It's often difficult to label all of the various aspects of the transcript, in addition, to the industry-specific jargon. Understanding that, we needed to find some literature that could help aid us in the project.

We concentrated on three main approaches in our research, which were all based upon and influenced by the BERT and FinBERT models. The first model we used was our interpretation of the FinBERT model, using the Financial Phrase Bank dataset to train the model on top of the BERT transformer. Our second model used the first model as a base, with the additional trainer layer of our earning calls dataset. This was done in an effort to further fine-tune the model towards earning calls data. The third model, like the first model, used our interpretation of the FinBERT model, but instead of using the Financial Phrase Bank dataset for the training, we used our earning calls dataset to train the model.

Due to resource problems and time constraints, we used a small version of the BERT model - bert_en_uncased_L-2_H-128_A-2, which uses only 2 layers, 128 hidden units and 2 attention heads. Since the earning calls are not labeled, we had to figure out a labeling algorithm to use in the supervised models.

There were three different techniques for labeling the earning calls sentiments. First, we labeled the sentiment based on the difference in the share price between the day of the call and the following day. Second, we based the label on the average net price difference in the duration between the earning calls data and the following 7 days. Lastly, we based this label on the percent change in prices between the day of the call and 7 days after.

Results

The first method we tested was the bag-of-words model which resulted in a 98% positive classification which we realized was not a useful method. However, we did use this method in our exploratory data analysis and to better understand that language from the calls.

Based on the methods outlined above we were able to derive the predictions accuracy and F1 score illustrated below.

Test	PostCall Δ	PostCallNet Δ	FirstDay Δ
2	0.39	0.24	0.26
4	0.34	0.45	0.35
6	0.39	0.28	0.4

Table 1: F1 score for the three binary classification models performed on the labeling sentiment technique

Test	PostCall Δ	PostCallNet Δ	FirstDay Δ
2	0.55	0.36	0.27
4	0.45	0.45	0.36
6	0.55	0.45	0.45

Table 2: Accuracy score for the three binary classification models performed on the labeling sentiment technique

The PostCallChange, PostCallNetChange, and FirstDayChange algorithms had a small chance of getting a Neutral label. We can base this off of the likelihood that the share price the day of an earning call and post call has likely shifted based on normal market conditions. Based

on that assumption we can ignore Tests 1, 3, and 5 which use Positive, Negative, and Neutral classification, while Tests 2, 4 and 6 classify only Positive and Negative sentiment.

The table above shows the PostCallChange algorithm which yields the best results, while Test 4 appears to give the best model selection. This model uses a binary classification (Positive and Negative) on a BERT base model which uses the Financial Phrase Bank dataset to train the model.

Analysis and Interpretation

By using natural language processing we weren't able to definitively determine if there was a strong correlation between the earnings call sentiment and the price change pre/post earnings call. As we can see from the results the best model used base BERT in addition to the Financial Phrase Bank dataset which enabled the model to perform better by using the financial news headings.

The first model we used was BERT and the Financial Phrase Bank to predict the earnings call. We can look at this method from the perspective of an unsupervised learning model. One caveat is that we did not train the model on the earnings and the outcome of that was it only enabled one class throughout the prediction process. The next model we used BERT as the base and trained the model using the earning calls dataset. There were some inconsistencies with the results which returned a low accuracy prediction. We concluded that some of the reasons for this low accuracy and F1 score was:

1. We used a very small version of the BERT model which only consists of 2 layers, 128 hidden units, and 2 attention heads. We did not use the full resources available for the model.

2. Using only 100 earning calls scripts, our training sample size is very small and likely not enough information to properly train the model.

Conclusions

Our interpretation of the results that we derived from our MVP is that our model could be improved upon. Based on our preliminary research we did not find any conclusive evidence that a (positive, negative, or neutral) sentiment does not strongly correlate with the pre and post-share price. We also found that many of the earning calls we analyzed had a positive sentiment, however, they may have missed key financial metrics for that quarter. This could be a direct result that many of these calls are scripted ahead of time and many of these companies understand the importance of the message that they portray to their shareholders.

Future Work

The aspect of this research that we'd like to continue is compiling more transcripts to help improve the accuracy of our model. We were limited to the number of transcripts that we could gather due to our time constraints but would like to see a comprehensive database that could archive this information to utilize in future model runs. Additionally, we could see our model include more performance metrics beyond just the share price. This could include things like Gross(Net) profit margin, Earnings per share, Current Ratio, and many more. The more variables we can add will help identify more aspects of the business beyond just the one metric we analyzed in our research. As we outlined in our conclusion we understand that many of the transcripts are scripted, to combat that, we could look at only reviewing the Question & Answers section to get a more authentic analysis.

References

Araci, Dogu. “FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models.” (June 2019). https://arxiv.org/pdf/1908.10063.pdf.

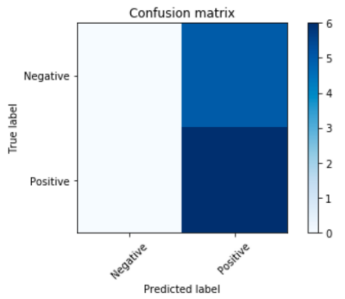
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” (May 2019). https://arxiv.org/pdf/1810.04805.pdf.

Howard, Jeremy, and Sebastian Ruder. “Universal Language Model Fine-tuning for Text Classification.” (May 2018). https://arxiv.org/pdf/1801.06146.pdf.

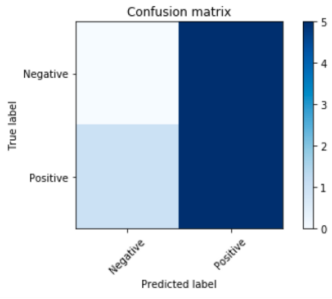
Mckay Price, S., James S. Doran, David R. Peterson, and Barbara A. Bliss. “Earnings conference calls and stock returns: The incremental informativeness of textual tone.” *Journal of Banking & Finance* 36, no. 4 (April 2012): 992-1011. doi.org/10.1016/j.jbankfin.2011.10.013.

Appendix 1

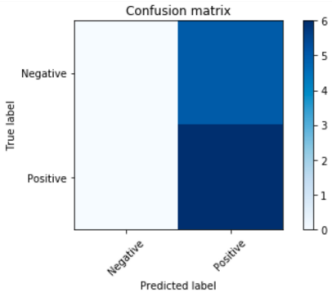
Confusion Matrices For the Different Models



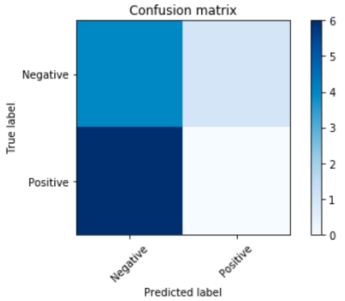
Model 1 using PostCallPctChange sentiment technique



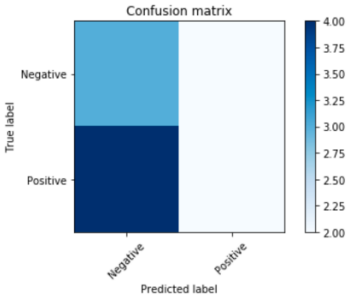
Model 2 using PostCallPctChange sentiment technique



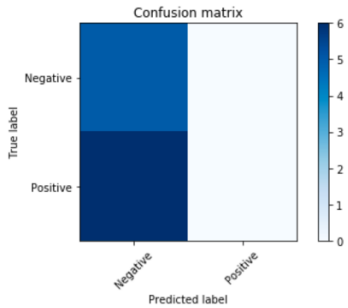
Model 3 using PostCallPctChange sentiment technique



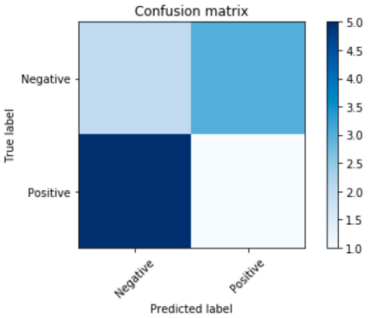
Model 1 using PostPostCallNetChangeCallNetChange sentiment technique



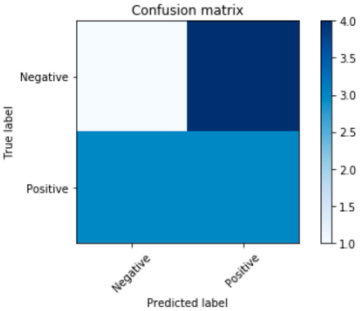
Model 2 using PostPostCallNetChangeCallNetChange sentiment technique



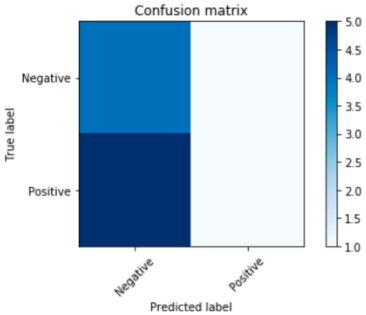
Model 3 using PostPostCallNetChangeCallNetChange sentiment technique



Model 1 using FirstDayChange sentiment technique



Model 2 using FirstDayChange sentiment technique



Model 3 using FirstDayChange sentiment technique