FinBERT Financial Sentiment Analysis Study

Author - Manan Patel
mbp001@ucsd.edu

1 Introduction

1.1 Task and Model Summary

This paper looks at the NLP task of Sentiment Analysis, specifically within the domain of Finance. The task involves taking a look at a variety of financial information sources such as news articles, financial statements of a company, and social media to predict the sentiment associated with a given security. The sentiment is classified into three categories: positive, negative, and neutral. This classification provides insights into public perception of a particular security and helps determine whether investors are likely to buy, hold, or sell that security. Financial sentiment analysis is thus crucial in forecasting the direction a security might take which is why this problem is so important.

In this Analysis Study, I evaluate **FinBERT**, a language model based on BERT and fine-tuned for NLP tasks in the financial domain. FinBERT was chosen due to its specialized fine tuning on financial texts, which allows the model to understand the finance specific terminology that are present in financial texts better than general-purpose language models. Previous studies have demonstrated how FinBERT's performance in financial sentiment analysis exceeds the performance of other models, making it a suitable choice for this analysis study [1].

1.2 Approach and Findings

To analyze FinBERT and identify its limitations, two datasets were used, each containing various indicators and metrics. One dataset focused on tweets regarding various securities, while the other consisted of financial news articles, with the title and summary of each article as the feature input. From the given text feature input of each dataset, I calculated various metrics to assess the model's performance on different types of data. These metrics include text length, the presence of the ticker symbol in the text, the number of hashtags, etc. By generating these different metrics, I could evaluate how FinBERT performed on diverse types of financial information, whether based on the source (Twitter or news) or specific features of the text (length, inclusion of security name, etc.). To conduct an in-depth analysis and detect the limitations of this model, I used ZenoML. This involved preprocessing the data and the output, then utilizing API calls to pass input data and receive predictions from the FinBERT model.

Two major limitations of FinBERT were revealed in this analysis study. First, FinBERT performed poorly on financial texts that were longer in length and those that had a formal, complex structure because the model could only support a maximum input size of 512 tokens. Thus, it was difficult for the model to learn long range dependencies seen in news articles due to the truncated output. The second limitation occurred because FinBERT was influenced greatly by keywords without looking at the broader context of the financial text. This is due to the fact that FinBERT is a fine-tuned model that extends from BERT. Because BERT was trained on a larger universal corpus of data where it learned common connotations of key words, FinBERT would utilize those learnt representation in contexts where it was incorrect.

2 Your Dataset

2.1 Data Source and Description

For this Analysis Study, I utilized two datasets from Hugging Face to determine the limitations of FinBert on different types of financial texts.

1. **Twitter Financial News Sentiment**
2. **Financial News Sentiment**

The first dataset, Twitter Financial News Sentiment, contains financial tweets that are labeled either 0 (negative), 1 (positive), or 2 (neutral). The dataset contains a total of **11,932** records. The tweets ranges from a string length of **2** to a maximum length of **190** with the url of the source of tweet at the very end of the input feature. **15.1%** of the dataset has bearish tweets, **20.2%** of the dataset has bullish tweets, and **64.7%** of the dataset has neutral tweets. An example record of this dataset is shown in **Table 1** below.

Twitter Tweet	**Label**
$BYND - JPMorgan reels in expectations on Beyond Meat https://t.co/bd0xbFGjkT	0

Table 1: Example record from the Twitter Financial News Sentiment dataset

The second dataset, Financial News Sentiment, contains news articles with three types of texts: summary of the article, title, and both summary and title of the article. For the purpose of this analysis study, I will utilize the input feature with both the summary and title of the article. The title and summary input feature ranges from a minimum string length of **64** to a maximum string length of **2,530**. The texts are classified, similar to the previous dataset, in three categories: 0 (negative), 1 (positive), and 2 (neutral). In addition to sentiment classification, this dataset also contains **10** topic labels that will be useful in understanding the limitations of FinBERT. There is a total of **1,510** records in this dataset. **3.2%** of the records in this dataset have a negative sentiment, **59.5%** of the dataset have a positive sentiment, and **37.2%** of the dataset have a neutral sentiment. An example record of this dataset is shown in **Table 2** below.

2.2 Data Preprocessing

For the Twitter Financial News Sentiment Dataset, the input feature contained the url of the source of the tweet at the very end of the string. This is a problem because the url is extraneous information in determining the actual sentiment of the financial tweet. Thus, I processed each tweet and removed the url at the end. Given that I was utilizing API calls to pass input test data into the FinBERT Model, there were limitations to the size of the data that could be fed to the model. Thus I restricted my input to the first **1000** records as the dataset was already randomly sorted. In addition, I calculate various metrics, discussed in depth later in **Section 3**, that would be later used as methods to analyze the limitations of FinBERT.

For the Financial News Sentiment Dataset, the data preprocessing was slightly different. Similar to the Financial News Sentiment, I restricted my input to the first **1000** records in order to accommodate for the limitation of the API calls to the FinBERT model. The FinBERT model API calls only allowed an input feature to have a maximum token length of **512**. Thus, I utilized the **AutoTokenizer** from the transformer library that was specifically pre-trained for FinBERT. Before inputting the piece of text into the model I would encode the text to get the number of tokens, restrict to under 512, and then decode the tokens to get the processed input. After that, I would normally pass the input to the FinBERT model through the API calls provided. In addition, I calculated various metrics from the input feature, similar to the first dataset, to further analyze the limitation of the model. For both datasets, I had to add specific columns, like an id column and correctness column, in order to upload the input feature and the output from the model to ZenoML.

3 Analysis Approach

To analyze FinBERT, I tested the ability of the model to correctly classify the sentiment of an input financial text based on various different features of the that particular text as described below. By utilizing the accuracy as a metric to determine the effectiveness of different input feature, I determined the limitations of FinBERT on specific features of financial texts.

3.1 Financial News

Financial news is an interesting metric to analyze FinBERT on due to several reasons. Unlike other sources of financial information, financial news articles are most likely created after a lot of research on the industry or security of topic. Moreover, the language of financial news articles may be neutral but can still indicate some strong sentiment in some direction. Thus it would be interesting to see how FinBERT performs on these type of text that are prone to be more formal and for a audience that has greater domain knowledge in finance.

Title and Summary of News Article	Label	Topic
Fortuna reports production of 101,840 gold equivalent ounces for the third quarter of 2022 – Fortuna Silver Mines Inc. (NYSE: FSM) (TSX: FVI) reports solid production results for the third quarter of 2022 from its four operating mines in the Americas and West Africa.	2	Quarterly financial Release

Table 2: Example record from the Financial News Sentiment dataset

3.2 Social Media Tweets

In contrast to Financial News, social media, such as twitter tweets, can sometimes come from reputable sources and other times for non reputable sources. It is more likely to be the case that social media tweets are informal and thus may or may not have greater sarcasm and exaggeration present in the tone of the text. The audience of social media tweets are for the greater public. This metric would be a good way to compare with Financial News to see how FinBERT performs on both formal and informal financial texts.

3.3 Financial Text Features

Texts in general come in varying forms. For both Financial News and Social Media Tweets, I parsed the input text and calculated several metrics in order to further analyze the limitations of FinBERT. Some of these metrics include string length, whether the text contains the name of ticker, whether it contains a date, whether it contains number, the number of hashtags, and the topic of the particular text. These specific features reveal interesting things about the financial texts. For example, the number of hashtags present can impact the tone/intent of the author of the tweet. By taking a look at these specific features of the text, I can gain a better understanding of how FinBERT learns representations of the text which can influence its ability to analyze sentiment.

4 Errors and their Categorization

Below is the high-level accuracy scores for both the datasets used in this analysis study. FinBERT was able to correctly classify the sentiment of the input financial texts from the Twitter Financial News Sentiment Dataset with an accuracy of **73%**. For the Financial News Sentiment Dataset, FinBERT had an accuracy of **37%**. **Figure 1** highlights how accurate FinBERT was able to classify each sentiment (negative, positive, neutral) across both the datasets.

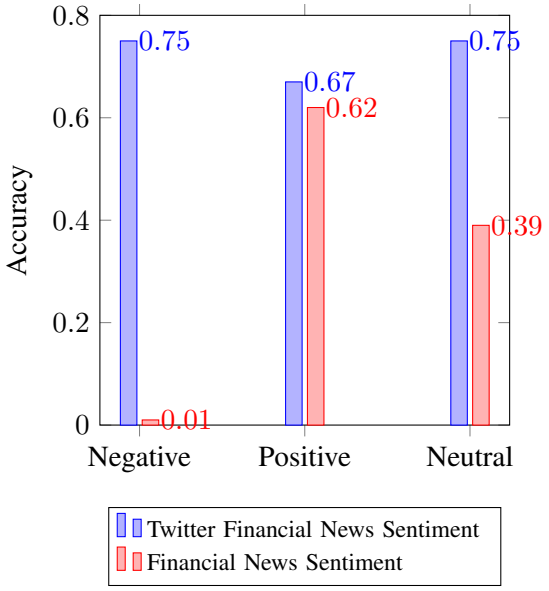

Figure 1: Comparison of FinBERT accuracy on different Sentiment Classes

4.1 Financial Text's Tone and Style

FinBERT performed much worse on the Financial News Sentiment Dataset with an accuracy of 37% as compared to the 73% accuracy on the Twitter Financial News Sentiment Dataset. **Table 3** highlights an example where the model predicted the sentiment of a positively labeled news article and tweet. Key things to note is that the news article is inherently longer in length as compared to the twitter tweet. Moreover the news article is packed with much more information and is structured more formally as compared to the tweet which is simple and straightforward. The accuracy scores in **Figure 1**, indicate a major limitation of FinBERT in its ability to process and classify texts that are structured formally and longer in length such as news articles.

The misclassified news article in **Table 3** highlights this limitation of FinBERT as described above. The article summarizes Alaris Equity Partner's Third Quarter Results, highlighting some clearly positive metrics such as its new investment

	Text Input	Label	Prediction
News	Alaris Equity Partners Income Trust Releases 2022 Third Quarter Results, Announces an Investment in Sagamore and a Unitholders' Distribution Increase of 3.0% – NOT FOR DISTRIBUTION IN THE UNITED STATES.	positive	negative
Tweet	$ABEO as expected, keeps going higher. Cantor doubled its price target this morning to $4	positive	positive

Table 3: Example predictions to model performance on different tones and styles of text

in Sagamore and a Unitholders' Distribution Increase of 3.0%. Later there is a clarification statement in all caps stating that this distribution is not for the United States. Out of context, this clarifying statement seems negative. The structure of this text is clearly formal as it tries to convey as much as information to the audience with clarification. When passing this input to the FinBERT model, it classifies this piece of financial information as negative. This reveals FinBERT's limitation to process and classify financial information that is not only longer in length with more complex context, but also classify texts that are written in a formal, structured style such as those with clarification statements.

4.2 Keyword-Based Decision-Making

When tested on both datasets, FinBERT would consistently error in classifying the sentiment of tweets and articles by utilizing key words to predict the sentiment. Below is an example of a tweet that was labeled as **neutral**, but instead was predicted by the FinBERT model to be **positive**.

`$AAPL - Max out Apple's Mac Pro for $52,599`

The tweet above simply states the price of a Apple's Mackbook given the best specification combination. Overall this tweet gives a neutral tone, however, the model predicts it to be positive. **Max** coupled with the cost $52,599 are the only words in the tweet above that may have a positive tone in certain context. FinBERT most likely takes the word Max out of context. Instead of interpreting it in the larger context of the tweet as describing the cost of a product, it prioritizes that word and predicts the tweet with a lack of context. Another example tweet where FinBERT prioritzes key words is shown below.

`Trans Mountain Costs to Increase 70% to $9.5 Billion, CBC Says`

This tweet was labeled in the Twitter Financial News Sentiment Dataset as negative, however FinBERT classified the sentiment of this tweet as positive. Similar to the above case the only word that has a positive tone is Increase. Looking at the context of the tweet as a whole, increase refers to costs which indicates a negative sentiment. However the model is not able to look at this broader context and prioritizes the word increase. This is most likely why FinBERT predicts this tweet as positive.

This keyword-based decision making error was not only consistent amongst tweets, but also with the news articles that were fed into the model. Below is just a portion of a news article summary that was inputted into FinBERT.

`a leading US-based franchise system providing drug testing, alcohol screening, DNA and clinical lab testing services announced today that it expects to release financial results for the third quarter of 2022 before the market open Monday, November 28, 2022.`

The news article above was labeled as positive in the Financial News Sentiment dataset, however the model classified the sentiment of this article as being negative. The input text summarizes that a particular US-based company is expecting to release its financial results. The only words that have a negative connotations are the words used to describe the service the company provides: drug testing, alcohol screening, etc. FinBERT most likely classifies this article with a negative sentiment because it takes these words such as drug testing, and alcohol screening out of context. Instead of viewing it in context of a company description, it prioritizes these key words independently which influences how the model classifies this text. This was consistent in various other cases such as the one

above.

Overall, it is clear that FinBERT incorrectly classifies an input financial text, no matter the style or length of text, by ignoring the entire context of the input and making decisions based on the common sentiment of certain key words.

5 Discussion

FinBERT is fine-tuned language model that was based off BERT. Thus, to understand the reasons for why FinBERT has the limitations described in **Section 4**, examining the architecture of BERT is necessary as shown in **Figure 2** below.

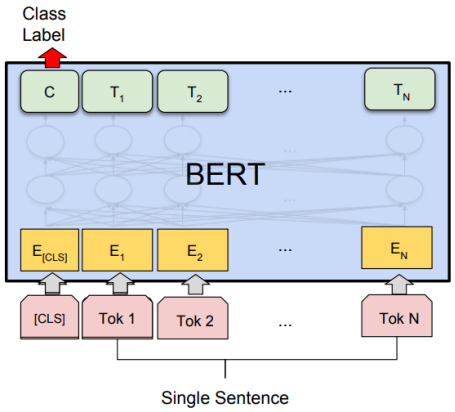

Figure 2: BERT architecture. Source:[3]

5.1 Financial Text's Tone and Style

One of the major limitations of FinBERT, as described in Section 4.1, is that it struggles to accurately classify financial texts that are longer in length and have a formal, complex structure.

BERT uses bidirectional context to allow it to understand the attention of a token both from the left and right direction. The only limitation with the BERT architecture, and as a result with the FinBERT architecture, is that the model's maximum input size is **512** tokens [2]. All the data that has this formal, complex structure came from the Financial News Sentiment Dataset and exceeded the maximum number of tokens allowed when decoded. As described in Section 2.2, Data Preprocessing, to utilize the model, I had to limit the input to under 512 tokens because the input data in this particular dataset exceeded the maximum capacity.

This limitation can explain why FinBERT struggled to understand the long-range dependencies of news articles because important context and clarifying statements might be truncated. This would lead to a loss of crucial information necessary for the model to accurately predict the sentiment of the input text. This truncation most likely affects the model's performance on texts that are more formal and complex, as these texts usually contain interdependent clauses and detailed clarifications that require a comprehensive understanding of the text.

5.2 Keyword-Based Decision-Making

Another major limitation of FinBERT in performing sentiment analysis on financial texts is that the model tends to prioritize keywords and often ignores the broader context of the entire text. As a result, the model's decision on which sentiment it predicts is largely influenced by the common connotations and sentiment associated with these keywords.

BERT utilizes subword tokenizations, like FinBERT, to represent each token. Although FinBERT is fine-tuned on top of BERT in order to understand specific terminology and nuances in financial texts, the initial pre-training done on BERT still may have a large influence on the model's representation of certain words. BERT was most likely trained on a large corpus of data and thus it would learn common connotations of several key words. As a result, these learned representations may still have a large influence when FinBERT performs sentiment analysis on financial texts.

6 The way forward (optional)

Modifying the architecture of FinBERT as well as improving the contextual representation that the model creates will improve the performance of FinBERT in sentiment analysis of financial texts.

Architecture Modification: Currently, FinBERT has a maximum input size of only 512 tokens. This is a problem because it prevents the model from being able to properly learn the long range dependencies seen in financial texts like news articles which can have more than 512 tokens. Thus modifying the architecture such that it can support input sizes greater than 512 tokens by either implementing hierarchical or chunk-based approaches to handle longer documents while retaining contextual understanding will solve this issue.

Contextual Embedding Refinement: To address the issue of FinBERT being influenced by BERT's pre-trained biases towards certain keywords, I would further fine-tune the model's contextual embeddings. This refinement aims to reduce the model's reliance on isolated keywords and improve its ability to capture the broader context of the text. One potential solution is to completely remove the learned representations from BERT and pre-train FinBERT specifically on financial texts so that doesn't get influenced by common key words.

7 Your Implementation (optional)

8 Conclusion

In the course of this analysis study, two significant limitations of FinBERT have come to light. Firstly, FinBERT's performance was extremely poor, accuracy of **37%**, when applied to lengthy financial texts with formal, intricate structures. This was mainly due to the fact that the FinBERT model could only have a maximum input size of 512 tokens. As a result, the model was unable to understand the extensive contextual and long range dependencies typical of news articles, as truncation led to vital information being overlooked.

The second limitation was due to FinBERT's tendency to heavily prioritize keywords, often at the cost of ignoring the broader context of the financial text. This behavior is most likely due to the fact that FinBERT was fine-tuned on BERT, which was originally trained on a comprehensive universal corpus. As a result, FinBERT inherits BERT's learnt understanding of common connotations associated with keywords. Despite being fine-tuned itself, it may be possible that BERT's learnt representations still have a major influence when the model classifies the sentiment of any financial texts.

Determining these limitations of FinBERT are important not only for this model's ability to perform financial sentiment analysis, but also for any other fine-tuned model that utilizes representations derived from pre-trained language models like BERT. Understanding and addressing these limitations are crucial steps towards improving the generalizability, robustness, and reliability for fine-tuned models like FinBERT across various domains. In the context of financial analysis, this is extremely important because it allows us improve the model so that it can better forecast whether investors will buy, hold, or sell a security.

References

[1] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] GeeksforGeeks. Sentiment classification using bert. Website, 2023.