Leveraging NLP for Predicting Reporting Lags from Earnings Calls

Jian Wang (j955wang@berkeley.edu) Jing Wen (jingwencrystal@berkeley.edu)

December 11, 2023

Abstract

In this pioneering study, we explored the relationship between earnings release conference calls and the reporting lag – the duration it takes for a company to submit its financial reports. Our methodology was comprehensive: we utilized BERT for analyzing raw transcripts, and augmented our analysis with audio data, GPT-generated summaries, and financial sentiment analysis. The results indicated a subtle but significant correlation between earnings calls and reporting lags, achieving a notable high accuracy of 58.77%, surpassing the baseline of 50.21%. Both BERT and GPT summaries played a key role in attaining this level of accuracy. Nevertheless, our dataset's inherent noise posed a limitation, hindering any further improvement in accuracy.

Unexpectedly, merging raw text with audio or sentiment analysis did not lead to better outcomes, which could be a consequence of the dataset's noise or BERT's relative efficacy over BiLSTM in this context. In scenarios involving a mix of textual and non-textual data, we opted for BiLSTM over BERT. Our findings suggest that a dataset with less noise might improve the accuracy of models that integrate raw text with audio and sentiment analysis, potentially yielding deeper insights in subsequent research endeavors.

1 Introduction and Background

The evolution of Natural Language Processing (NLP), especially with the emergence of BERT (Bidirectional Encoder Representations from Transformers), has significantly transformed the finance sector [1]. In this industry, the analysis of extensive textual data, including market reports, financial news, and regulatory filings, is essential. This research report serves as exploratory work to prove that effective usage of advanced NLP techniques could substantially predict some of the important metrics in the financial market.

A key element of finance reports, the reporting lag — the duration required for a company to submit its financial reports — is notably one of the most important factors to reflect the reporting speed of a company, and act as an key indicator of operational efficiency and transparency. Previous studies have investigated the influence of Reporting Lag on stock performance [2]. Observations indicate that companies with quicker reporting times generally experience a market premium compared to those with slower reporting. Significant findings demonstrate that both characteristics specific to the firm and attributes of the documents play a crucial role in the length of Reporting Lags, with shorter lags often correlating with positive earnings surprises. Conversely, longer Reporting Lags might suggest management's intentional delay in disclosing unfavorable news.

Other research has shown the viability of using verbal and textual cues from earnings calls to forecast financial risks [3]. This research encompassed data from 280 companies in 2017, with each company potentially presenting 1 to 4 earnings calls. For each call, transcripts and audio recordings were analyzed. A bidirectional LSTM framework was utilized, yielding higher prediction accuracy. At the time, BERT was not publicly available, so sentence-level embeddings were used instead.

Our research is designed to explore the potential of quarterly earnings call data in forecasting Reporting Lags for the respective fiscal year. We intend to harness cutting-edge NLP methodologies, with a specific focus on Transformer models and BERT, to assess their efficacy in providing a more nuanced understanding of Reporting Lags through the detailed content of earnings calls. Furthermore, our study will integrate and analyze speaker audio data alongside BERT to enrich the analysis. This audio data, a crucial component of our dataset, offers an additional dimension for interpreting the calls. We will also add the sentiment analysis to accompany the transcript text. We will employ the renowned Loughran-McDonald Sentiment Word Lists. These lists are specially tailored for financial text analysis, thereby augmenting our analytical framework with deeper, sector-specific insights.

2 Dataset

2.1 Source

Our research utilizes a pre-existing dataset comprising transcripts and audio recordings from the 2017 CEO's earnings calls of 280 companies. It includes approximately 700 SP earnings conference calls, consisting of 88,829 sentences.This dataset serves as the primary source for our analysis.

For calculating the reporting lag, we employed the Python package sec-cik-mapper in conjunction with data from the SEC's official EDGAR website. The sec-cik-mapper package assists in creating correspondences between stock and mutual fund identifiers as per SEC standards. Our data retrieval process involved sending requests, structured based on the website's JavaScript requests, to acquire responses containing key information such as "file_date" and "period_ending". The reporting lag was computed as the number of business days between the "file_date" and "period_ending".

We encountered challenges with inconsistencies in company names within the dataset, which sometimes led to difficulties in retrieving data from the EDGAR website. In instances where a company's name in the dataset did not yield results, we manually identified the corresponding company by examining the earnings call transcript. This manual approach also became necessary when the automated data retrieval process inadvertently fetched the filing date of an amended 10-K form instead of the original 10-K, resulting in an erroneously extended reporting lag. Such instances were identified and corrected through manual verification.

2.2 Noise

Another complexity we faced was the inconsistency in the number of earnings releases corresponding to a single reporting lag within the dataset. Ideally, four earnings releases should align with one reporting lag. However, our dataset often presented fewer earnings releases per reporting lag, introducing noise and limiting the precision of our analysis. Time constraints hindered our ability to acquire additional transcripts and audio files for the missing earnings releases, affecting the overall accuracy of our dataset.

2.3 Initial Data Exploration

To explore the dataset, we established a basic neural network as our initial predictive model and employed a state-of-the-art pretrained sentence-level transformer known as ”all-MinLM-L6-v2”. This transformer model is capable of generating sentence-level embeddings for any input text. To simplify the process, we treated the entire earnings call data as a single sentence, utilizing the resulting 346-dimensional vectors as input features for our predictive model. These features were then fed into a straightforward fully-connected neural network for standard regression tasks, utilizing Mean Squared Error (MSE) as the basis for the loss function.

Regrettably, this model exhibited suboptimal performance, achieving only a 0.06 $R\hat{2}$ score in regression tasks. This outcome can be attributed to several factors:

- **Model** Sentence-level transformers may not effectively capture crucial information from extensive textual data.

- **Data** The size of our datasets is insufficient for learning meaningful parameters in a regression-based task. Predicting exact reporting lags based solely on earnings conference call information proves challenging. To enhance accuracy, it becomes necessary to incorporate country and industry-specific adjustments to the reporting lags and focus on predicting relative changes rather than absolute values.

As shown in the Appendix Figure 1, although our resulting prediction is relatively linear to the actual reporting lag, the accuracy of the prediction is extremely low. our prediction ranges from 24 to 28 days, while the actual reporting lag vary from 20 up to 70 days.

This practices prompted a shift in our approach towards transforming our tasks into classification tasks rather than precise predictions.

2.4 Y Labels

Consequently, we opted for a binary classification approach for our dependent variable, ’y’. This variable is labeled ’1’ for instances where the company’s reporting lag exceeds the median value of 37, signifying a longer-than-average reporting duration. Conversely, cases where the reporting lag is less than or equal to the median are labeled ’0’. This binary classification framework allows for a more streamlined and focused analysis within the bounds of our dataset’s limitations.

3 Research Method

3.1 Baseline

In our methodology, we generated random binary values (0 or 1) and compared these with the ’y’ labels in our dataset. To establish a reliable baseline, this random selection and comparison process was iteratively conducted 20 times. The average accuracy from these iterations was calculated, resulting in a baseline figure of 50.21%.

3.2 Methodology

In our study, we utilized four distinct types of inputs: raw text, summaries generated by GPT, audio features, and sentiment analysis features. We initially processed each input type independently. Subsequently, we experimented with combinations of two different inputs. Since BERT is adept at handling tokenized text inputs, we directly fed raw text and GPT-generated summaries into BERT. However, as the audio features and sentiment analysis features are numerical, they were not suitable for direct processing with BERT. For any model that incorporated these numerical inputs alongside text, we tokenized the text and utilized a BiLSTM approach for processing the combined inputs. This methodology allowed us to leverage the strengths of both BERT for text-based data and BiLSTM for mixed data types.

3.3 Models

3.3.1 BERT

In our research study, we selected a maximum sequence length of 128 tokens for the BERT model, primarily due to memory constraints encountered with longer sequences. Attempts to process longer texts by dividing them into smaller segments proved to be excessively time-consuming for each file, rendering the task impractical given our computational resources. Fortuitously, we observed that the initial segments of the conference calls typically contain the most critical information. Additionally, we experimented with various learning rates and dropout rates, ultimately selecting the most effective combination identified in our tests: a learning rate of 0.00005 and a dropout rate of 0.3.

3.3.2 LLM Summarization and Prompt Engineering

It was highlighted that a typical earnings call comprises more than 4000 tokens, posing a considerable challenge for NLP models such as BERT and transformers to glean valuable insights, particularly when dealing with relatively small datasets. To address this, it becomes imperative to undertake dimensional reduction and summarization, focusing on extracting pertinent information from the earnings call. In this context, we employed cutting-edge tricks —prompt engineering—and delegated the summarization task to the advanced large language model 'GPT-3.5-turbo-1106.', developed by openAI.

In our approach, we input the entire textual information of each earnings call into the GPT-3.5 framework. A standard prompt is provided at the beginning of the text, requesting the model to summarize the tone of the earnings call and evaluate the CEO's confidence in the earnings results. The expected response includes a general summarization of the entire earnings call and an internal rating reflecting the document's strength.

For instance, in the case of the 2017 Q2 earnings call for 'Advanced Micro Devices Inc,' which originally contained over 4500 tokens, the GPT-3.5 turbo-generated summarization reads: 'The tone of the earnings call is strong, with the CEO expressing confidence in the company's performance and product portfolio. The CEO addresses potential concerns by noting an expected increase in the ramp of game consoles in the second quarter. Overall, the strength of the earnings call is rated as 4.'

3.3.3 Sentiment Analysis

We employed the Loughran-McDonald Sentiment Word Lists for sentiment analysis, leveraging their specialized design for financial text analysis. Tim Loughran and Bill McDonald's creation of these lists marks a notable development in evaluating sentiments within financial documents, such as earnings reports. These lists adeptly categorize words into distinct sentiment categories, including positive, negative, uncertainty, litigious, strong modal, and weak modal. This categorization is particularly useful in financial contexts, where the sentiment connotation of certain words can differ significantly from general usage. For example, a term like "liability" often carries a negative sentiment in financial documents, whereas it might be perceived as neutral or even positive in general discourse [4]. This approach allowed us to integrate a sophisticated sentiment analysis framework into our prediction model.

3.3.4 Audio Analysis

The highlight of our research is to combine both audio and textual information into the prediction process. If the textual presentation given by the CEO matters and can have prediction power on transparency of the company's future, then the way they talk should matter as well. We utilized the librosa library for audio analysis, driven by the premise that the nuances in speech patterns are significant. Given the constraints imposed by the size of our datasets, we made a strategic decision to create aggregate features for each earnings call rather than generating multiple features at the sentence level. For every sentence, we derived 14 frequency-related features, including zero-crossing rate, tempo rate, average speaking speed, and frequency, as well as standard deviation of speaking speed. Subsequently, we computed the average of these sentence-level audio features with the aim of obtaining generalized insights into the manner in which people speak. The objective of these audio features is to capture variations in speech patterns.

In our architectural design, we integrate output features from both textual neural networks (BERT) and audio neural networks. These features are assembled into the final neural network, where we execute prediction tasks, encompassing both regressions and classifications. A visual representation of the general architecture is in the Appendix Figure 2.

4 Results

4.1 BERT

Our initial application of BERT, using raw text inputs, yielded an accuracy of 58.77%. We observed more consistent outcomes when the text was processed to exclude numbers. An alternative approach involved utilizing GPT-generated summaries as input for BERT, which interestingly also resulted in a similar accuracy of 58.77%. Further, we experimented with a combination of raw text and summary as inputs [raw text, summary], but encountered a 'longest_first' truncation strategy warning and again achieved an accuracy of 58.77%. This suggests that the summary portion may not have been effectively utilized, possibly due to the truncation strategy.

Pursuing a different strategy, we experimented with stacking the raw text and summary inputs instead of combining them. This effectively doubled the data volume. However, this approach led to a slightly lower accuracy of 57.46%, potentially due to memory constraints exacerbated by increased epochs. Detailed results are in the Appendix Table 1.

4.2 Audio and Sentiment Analysis

In our study, we worked with a dataset comprising 26 features for audio. Various combinations of layers, dropout rates, and learning rates were tested, yielding a maximum accuracy of 56.14%. We then experimented with dual-input models for the BiLSTM, combining Audio with summaries and Audio with raw text. These configurations resulted in slight accuracy decreases to 55.26% and 56.14%, respectively. Focusing on Sentiment Analysis, which includes six features (positive, negative, uncertainty, litigious, strong modal, weak modal), the highest standalone accuracy was 55.26%. When we combined Sentiment Analysis with summaries as inputs for the BiLSTM, there was an increase in accuracy to 57.89%. Conversely, merging Sentiment Analysis with raw text led to a reduced accuracy of 54.39%. A further combination of Audio and Sentiment Analysis input into the BiLSTM resulted in an accuracy of 46.49%.

Additionally, we explored a tri-input approach, integrating Summary with both Audio and Sentiment Analysis. This configuration yielded an accuracy of 58.77%. Detailed results are in the Appendix Table 2.

5 Discussion

After exploring various methodologies, our findings indicate that the highest achievable accuracy was 58.77%. Employing BERT to process either raw text or summaries consistently resulted in this level of accuracy. Our analysis indicates that BERT surpasses both Sentiment Analysis and Audio-based methods in accuracy. However, contrary to expectations and existing literature that highlights improved accuracy with audio features, combining Sentiment Analysis or Audio features with BERT did not yield any accuracy gains in our study. This discrepancy may stem from the fact that their dataset likely contained less noise and had a more direct correlation between features and labels. Our dataset, in contrast, seems to have an inherent accuracy limitation of 58.77%, a threshold we were unable to surpass.

We theorize that an integrated approach combining Sentiment Analysis and Audio features with either raw text or summaries could potentially elevate accuracy levels. Regrettably, due to constraints in memory capacity, we were unable to experimentally verify this hypothesis.

Our study did not achieve high accuracy, primarily due to two key factors. Firstly, the limitation posed by the small amount of available data was a significant constraint. More critically, however, was the issue of data quality: our dataset was marred by considerable noise. In an ideal scenario, all conference calls would occur before the filing date, with consistently toned discussions. Yet, this pattern was not universally observed. Take Parker Hannifin Corporation, for example, which was labeled as '1', indicating a filing later than the median. This company's filing date was 2017-06-30, preceded by three conference calls: the first in February (medium to strong tone), the second in April (strong tone), and the last in November (medium tone). Our model, constrained by time, labeled these entries as '1'. However, the situation is complicated by the existence of two other preceding conference calls not included in our dataset, which may have contributed to the late filing. Furthermore, the last conference call, occurring post the 10-K filing, should not have impacted the 'y' value.

6 Conclusion

Our study marks the inaugural exploration of the correlation between earnings release conference calls and reporting lag. The key findings of our research are as follows:

- A discernible relationship exists between earnings release conference calls and reporting lag. This is evidenced by the performance of our BERT model, which achieved an accuracy of 58.77%, surpassing our baseline accuracy of 50.21%.
- The integration of Text with Audio or Sentiment Analysis did not yield an increase in accuracy. This could be attributed either to the presence of noise in our data or to the more sophisticated capabilities of BERT in comparison to BiLSTM.
- There is a potential that augmenting our model with additional features, such as combining Text with Audio and Sentiment Analysis, might enhance accuracy. However, due to memory limitations, this hypothesis remains untested.

In conclusion, our research contributes significantly to the understanding of how to predict reporting lags in the finance sector, employing advanced NLP techniques. Despite encountering challenges and dealing with relatively noisy data, our approach offers valuable insights and lays a foundation for future advancements and detailed investigations in this dynamic area of study.

Appendix

Figures

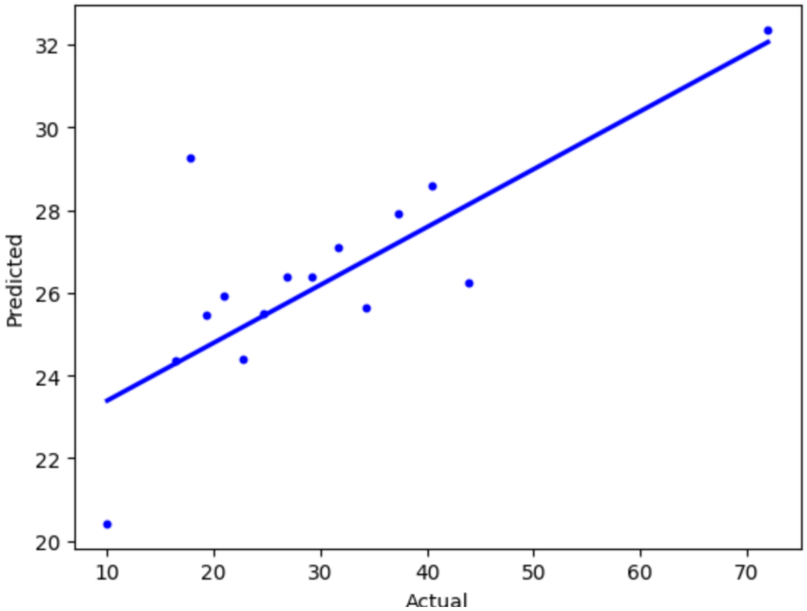


Figure 1: Linearity Predicted vs Actual

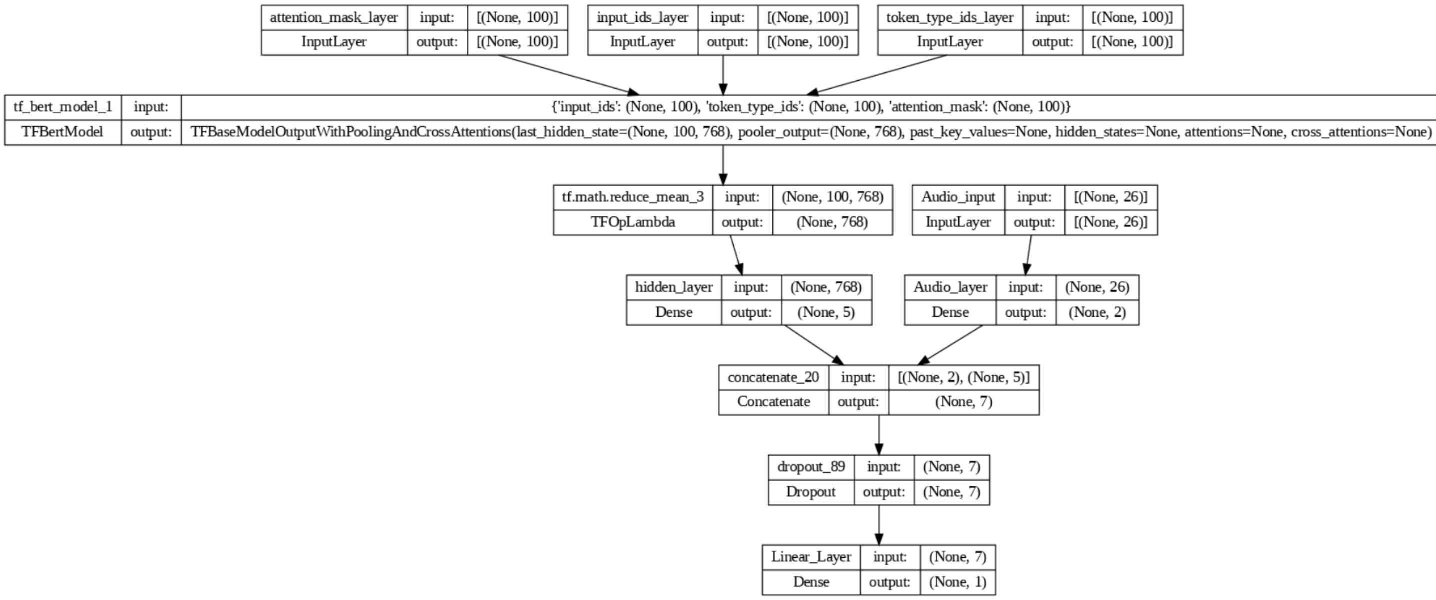


Figure 2: Model Architecture

Tables

Inputs	Model	Optimal
Raw Text	Bert	58.77%
Raw Text (no numbers)	Bert	58.77%
Summary	Bert	58.77%
[Raw Text, Summary]	Bert	58.77%
Raw Text + Summary	Bert	57.46%

Table 1: Bert Results

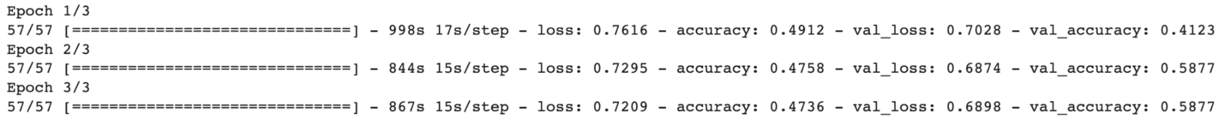
Inputs	Model	Optimal
Audio	BiLSTM	56.14%
[Audio, Summary]	BiLSTM	55.26%
[Audio, Raw Text]	BiLSTM	56.14%
Sentiment Analysis	BiLSTM	55.26%
[Sentiment Analysis, Summary]	BiLSTM	57.89%

Table 2: Audio and Sentiment Analysis Results

References

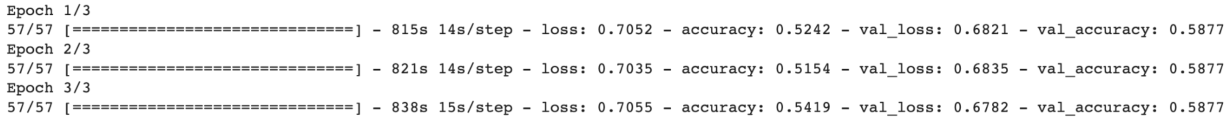
1. Gokhan, T., Smith, P., & Lee, M. (2021). Extractive financial narrative summarization using SentenceBERT-based clustering.
2. Bannouh, K., Geng, J., & Peeters, B. (2021, February). Filing, Fast and Slow: Reporting Lag and Stock Returns.
3. Qin, Y., & Yang, Y. (2019, February). What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1038.pdf
4. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66(1), 35-65.

Raw Outputs



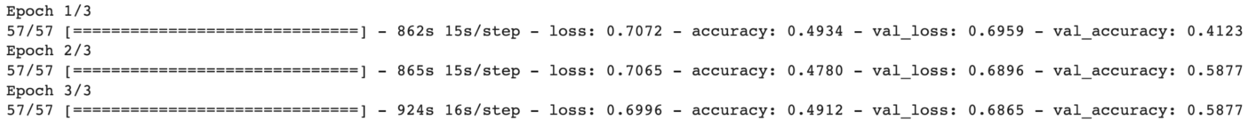
```
Epoch 1/3
57/57 [==============================] - 998s 17s/step - loss: 0.7616 - accuracy: 0.4912 - val_loss: 0.7028 - val_accuracy: 0.4123
Epoch 2/3
57/57 [==============================] - 844s 15s/step - loss: 0.7295 - accuracy: 0.4758 - val_loss: 0.6874 - val_accuracy: 0.5877
Epoch 3/3
57/57 [==============================] - 867s 15s/step - loss: 0.7209 - accuracy: 0.4736 - val_loss: 0.6898 - val_accuracy: 0.5877
```

Bert with numbers



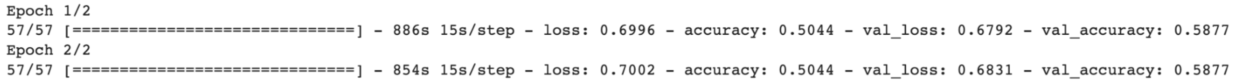
```
Epoch 1/3
57/57 [==============================] - 815s 14s/step - loss: 0.7052 - accuracy: 0.5242 - val_loss: 0.6821 - val_accuracy: 0.5877
Epoch 2/3
57/57 [==============================] - 821s 14s/step - loss: 0.7035 - accuracy: 0.5154 - val_loss: 0.6835 - val_accuracy: 0.5877
Epoch 3/3
57/57 [==============================] - 838s 15s/step - loss: 0.7055 - accuracy: 0.5419 - val_loss: 0.6782 - val_accuracy: 0.5877
```

Bert with no numbers



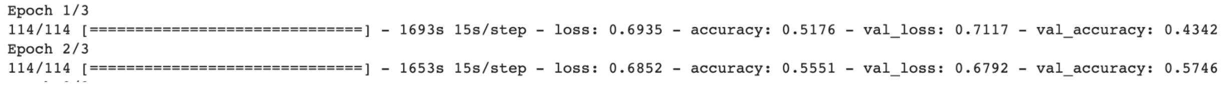
```
Epoch 1/3
57/57 [==============================] - 862s 15s/step - loss: 0.7072 - accuracy: 0.4934 - val_loss: 0.6959 - val_accuracy: 0.4123
Epoch 2/3
57/57 [==============================] - 865s 15s/step - loss: 0.7065 - accuracy: 0.4780 - val_loss: 0.6896 - val_accuracy: 0.5877
Epoch 3/3
57/57 [==============================] - 924s 16s/step - loss: 0.6996 - accuracy: 0.4912 - val_loss: 0.6865 - val_accuracy: 0.5877
```

Bert with summaries



```
Epoch 1/2
57/57 [==============================] - 886s 15s/step - loss: 0.6996 - accuracy: 0.5044 - val_loss: 0.6792 - val_accuracy: 0.5877
Epoch 2/2
57/57 [==============================] - 854s 15s/step - loss: 0.7002 - accuracy: 0.5044 - val_loss: 0.6831 - val_accuracy: 0.5877
```

Bert with raw text and summaries



```
Epoch 1/3
114/114 [==============================] - 1693s 15s/step - loss: 0.6935 - accuracy: 0.5176 - val_loss: 0.7117 - val_accuracy: 0.4342
Epoch 2/3
114/114 [==============================] - 1653s 15s/step - loss: 0.6852 - accuracy: 0.5551 - val_loss: 0.6792 - val_accuracy: 0.5746
```

Bert with stacked raw text and summaries