Using BERT to Enhance Multi-Classification of Finance-Related Tweets

Vincent Goldberg, Samantha Williams
University of California, Berkeley – School of Information
W266 – Natural Language Processing with Deep Learning
{vincent.goldberg, swllms} @berkeley.edu

August 2023

Abstract

In today's landscape, Twitter has become an indispensable platform for disseminating financial market insights, investment sentiments, and expressions of financial perspectives. This paper delves into the realm of classification models by comparing the efficacy of BERT and FinBERT in categorizing finance-related tweets in real time. To enhance precision, the investigation employs strategies such as cluster inter-training, fine-tuning, and token masking. The study sheds light on the intricate task of classifying financial tweets, a challenge arising from their distinctive language and content. Despite numerous attempts to refine accuracy, neither model emerged as the unequivocal leader, prompting the suggestion for future endeavors to amalgamate models for specific categories and to delve into larger datasets for more refined fine-tuning. Ultimately, this research contributes to an improved comprehension of financial trends and sentiments within the realm of social media discourse.

1 Introduction

Social media platforms have become the bedrock of disseminating information, opinions, and sentiments related to financial markets and investment activities. Twitter, with its sheer volume and rapid dissemination of tweets in real-time, has gained significant prominence as a platform where users openly express their views on various financial topics. Thus, it is essential to develop effective methods for categorizing and analyzing finance-related content in real time.

It is our intention to design and implement an accurate classification model that can efficiently categorize finance-related tweets based on their topics. To achieve this goal, state-of-the-art BERT models will be leveraged, aiming for enhanced accuracy in the classification process.

Although BERT has consistently achieved impressive results across various NLP tasks, our hypothesis is based on a dual-pronged approach aimed at achieving even higher classification accuracy. First, by leveraging the domain-specific pre-trained model FinBERT (Yang et al. 2020) that is tailored to the financial domain, we aim to capitalize on its domain-specific knowledge to potentially enhance tweet categorization performance. Second, we adopt well-established methodologies for fine-tuning transformer models (Sun et al., 2019) including cluster inter-training and parameter hyper tuning, as well as further fine-tuning through token masking (Shnarch et al. 2022). By combining these approaches, we endeavor to produce a highly accurate and efficient model for classifying finance-related tweets, thereby enabling better comprehension of financial trends and market sentiments.

For this paper, it is presumed that the reader has a proficient knowledge of NLP and a comprehensive understanding of BERT (Devlin et al., 2019). However, we will note relevant works used for research and framing of our models, evaluations and further experiments and will provide further context where appropriate to understanding our approach.

2 Data Collection and Preprocessing

The Twitter Financial News dataset from Kaggle consists of 21,107 annotated records with 20 distinct labels categorizing finance-related topics such as earnings, currencies, macro, and company news. The downloaded dataset contained a training set of 16,990 records and a validation set of 4,118 records. To assess our models' generalization, the validation set provided was treated as our test set. We created a new validation set from the training data for monitoring model performance during training. The final training set consisted of 13,592 records, the final validation set had 3,398, and the test set remained the same with 4,118 records.

Preprocessing the tweet texts was essential due to their unique characteristics as their content could contain hyperlinks in bitly format, emojis, and/or various special characters such as "&" and "$." During the preprocessing phase, we opted to remove only the hyperlinks because the bitly format held little contextual value. We retained the emojis and other symbols, as they may carry valuable sentiment information when used with a special tokenizer and are tolerated by the BERT models with the use of unknown tokens (Devlin et al., 2019).

The initial analysis showed that tweet lengths varied considerably, with the shortest tweet consisting of only two characters and the longest extending to 316 characters. However, the tweet length distribution followed a normal distribution pattern making it easier to identify outliers within the dataset. Examination of the label distribution in the dataset reveals an imbalance across the 20 labels. Rather than balancing the labels during the data split, we opted to keep the data unbalanced to mirror the real-world distribution of finance-related topics. This approach provides a more realistic evaluation of the model's performance, reflecting its behavior in practical applications. By leaving the data unbalanced, we aim to ensure that the model learns from the naturally occurring class distribution it is likely to encounter during real-world deployment.

3 Models

The experiments utilized two pre-trained transformers from the Hugging Face model hub, “bert-base-uncased” (Devlin et al., 2019) and “yiyanghkust/FinBERT-pretrain” (Yang et al., 2020). The pooled output of these models were fed into a Keras neural network that employed a single hidden layer followed by a dropout layer before reaching the final classification output layer. This structure and the following hyperparameters were used for all experimental versions: a hidden size of 201 to capture the complex patterns contained in the data set while balancing the memory and computation efficiency; a dropout rate of 30% to combat overfitting; a learning rate of 0.00005 so the model adapts gradually while preserving the knowledge of the pre-trained models; and a SoftMax activation function for classifying. During experimentation, the learning rate was the only hyperparameter to change when following the strategy of Sun et al. (2019). Finally, these models were compared to linear baselines as demonstrated by Lin et al., (2023).

3.1 Baseline

While we are embracing complex models, it is crucial not to overlook the effectiveness of humble linear classifiers as simple yet valuable baselines for text classification tasks as demonstrated by their competitive performance, interpretability, and efficiency. We followed the architecture used in Lin et al. (2023) which implemented a Linear SVM model for text classification. Following this strategy, we establish a strong baseline of 0.8438 accuracy with a macro F1 Score of 0.8248.

3.2 Base Models

Although BERT showcases impressive capabilities in general NLP tasks, its optimal performance within the financial domain is hindered by the presence of specialized jargon, distinct linguistic patterns, and context-specific meanings unique to the financial industry. To address this challenge, FinBERT (Yang et al. 2020) was pre-trained by analyzing financial communications, including corporate earnings reports, financial news articles, and other relevant financial documents. With a financial corpus consisting of 4.9 billion tokens and fine-tuning, FinBERT achieves a higher accuracy compared to BERT on three financial sentiment analysis tasks across 3 different datasets: Financial Phrase Bank, FiQA, and AnalystTone. As a result, we estimated it to have a higher accuracy in classifying our dataset.

3.3 Further Pre-Training (MLM)

Shnarch et al., (2022) propose further pre-training of the model using the self-supervised masked language model (MLM) task on unlabeled data from the target task domain as a strategy for enhancing performance. To implement this approach, we followed the outlined methods by preprocessing our training and validation sets, where individual tweets were stripped of their labels, concatenated, and then split into chunks of 100 tokens each. Finally, the MLM training was conducted, and the models were saved as checkpoints to be used later in the final classification models. We evaluated the progress of both BERTbase and FinBERT based on their perplexity scores before and after 5 epochs of training, resulting in significant reductions of more than 86% and 76%, respectively.

3.4 Cluster Inter-Training (CIT)

Because of the limited labeled data available and the dynamic nature of tweets, we find the introduction of new categories occurring in the real world. Standard classifiers often face challenges in handling these new categories as they lack specific information and context related to these novel classes. To combat this cold start, Shnarch et al., (2022) have shown performance improvements by adding an inter-training step, where BERT models are further pre-trained on the same data that's been clustered into further categories. We replicated this approach by implementing a sequential Information Bottleneck (sIB) with Bag of Words (BOG) representations to identify 50 different clusters in our data set. Both Bert models were trained, and the learned parameters were saved to be used later.

3.5 Fine-Tuning (FIT)

Sun et al., (2019) propose that aside from pre-training, classification can further improve through the fine-tuning of hyperparameters–specifically, by optimizing the learning rate. As such, the final approach undertaken during experimentation implemented a learning rate schedule. The schedule implemented a decay which incrementally decreased the initial learning rate of 0.00005 down to zero. This approach was applied to both base models and incorporated into our combination models right before the final classification.

3.6 Model Variations

A fundamental component of BERT models is their unique ability to learn and to transfer that learning to produce superior results (Devlin et al., 2019). To capitalize on this transfer learning, our experiments sought to combine the checkpoints of the previously individually run models. The approach taken was to order the checkpoints based on the type of training and followed a sequence of pretraining (MLM) first, then inter-training (CIT), and finally to add the learning rate fine-tuning as the final step (Shnarch et al., 2022). Multiple permutations were devised to experiment with the different combinations which are shown in Figure 1.

4 Results & Discussion

Table 1 provides a comprehensive overview of the

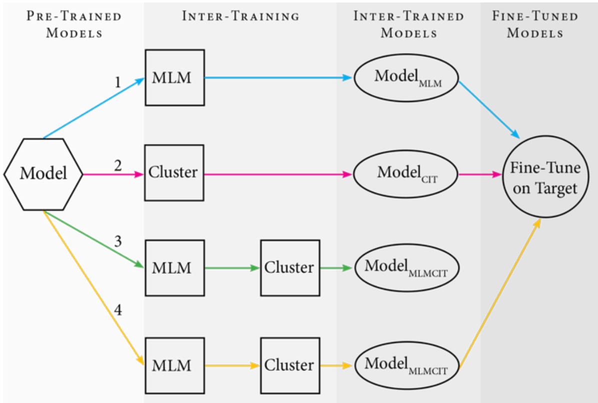

Figure 1: Diagram of the training combinations used for BERTbase and FinBERT

results obtained from various models employed in this study for categorizing finance-related tweets.

4.1 Evaluation Metrics

To evaluate the models, we relied on two metrics: accuracy and Macro F1 Score. However, due to the unbalanced nature of our data set, the primary score we use for comparison between models was the Macro F1 Score, which is what we reference in the following analysis. Finally, for evaluating the difference between the BERTbase and FinBERT models of our winning optimization strategy, we implement a delta confusion matrix to showcase what class each model struggled with predicting.

4.2 Model Results

Based on its domain-specific knowledge, we forecasted FinBERT would produce superior results when classifying the financial tweets. Indeed, during a one-to-one comparison of the vanilla models, FinBERT outperformed BERTbase. However, the results were small, creating roughly a 2% delta in Macro F1 Score. Despite this initial success, the following experimentation yielded surprising results. In all the implementations, BERTbase outperformed

FinBERT. The highest performing model was BERT-base-Fit with an accuracy score of 0.8834, improving nearly 2% from its initial evaluation. As such, BERT-base-Fit was the model used to evaluate the results.

4.3 Evaluating Results

The data sets used for classification in Sun et al., (2019) ranged from 54,000 to 120,000 records,

Table 1: Model Results		
Model	Accuracy	F1 Macro
Baseline (Scikit-Learn)		
Linear SVM	0.8438	0.8248
Base Models Uncased		
BERT-base	0.8664	0.8599
FinBERT	**0.8749**	**0.8795**
Fine-Tuning (Fit)		
BERT-base-Fit	**0.8880**	**0.8834**
FinBERT-Fit	0.8736	0.8757
Further Pre-Training (MLM)		
BERT-base-MLM	0.8868	0.8776
FinBERT-MLM	0.8713	0.8600
Cluster Inter-Training (Cit)		
BERT-base-Cit	0.8703	0.8566
FinBERT-Cit	0.8577	0.8494
Training Combos		
BERT-base-MLM-Fit	0.8786	0.8486
FinBERT-MLM-Fit	0.8679	0.8619
BERT-base-Cit-Fit	0.8679	0.8475
FinBERT-Cit-Fit	0.8713	0.8517
BERT-base-MLM-Cit	0.8747	0.8691
FinBERT-MLM-Cit	0.8650	0.8282
BERT-base-MLM-Cit-Fit	0.8829	0.8667
FinBERT-MLM-Cit-Fit	0.8594	0.8527

Table 1: Model Results for each model run with results for test accuracy and Macro F1 Score.

giving the models more opportunity to learn valuable insights needed from classification. Our training set consisted of 13,592 records, and we believed the small amount of data available for fine-tuning was impacting model performance. To test this theory, we reduced our training data by 50%, retrained and tested the model. What resulted was a decrease of 8% accuracy from 0.8834 to 0.8037. Although this is not drastic, we believe that this signaled that our limited data was limiting the models' ability to learn, thereby decreasing its ability to fully harness the power of further pre-training.

Next, we examined samples of the tweets our models were misclassifying. As the examples show in table 2, the predicted labels could easily be considered correct. In fact, when analyzing our results to see if the correct label was in the top three predictions of the model, the accuracy jumps up to 0.9689. Reviewing our data source, it is unclear by whom the labels were assigned and the strategy they used to develop the classifications. Nonetheless, our results indicate that using a multi-label classification structure would be the superior approach for tweet classification.

Finally, to see why FinBERT underperformed, we developed a confusion matrix showing (Figure 2) the deltas between the BERTbase and FinBERT predictions. The positive deltas show the areas where BERTbase excelled. BERTbase beats FinBERT in several categories one would think a financially oriented transformer would excel, such as “M&A | Investments”, “Stock Commentary”, “Financials”, and “Fed | Central Bank”. It appears that the highly technical and verbose format of the FinBERT corpora may inhibit its ability to properly categorize financial topics in a tweet format. Moveover, the broad generalization of the BERTbase model, combined with the fine-tuning on the task-specific data set appear to outperform

Table 2: Label Prediction Errors			
Model	Text	True	Predicted
BERTbase	Soybeans weak today but appears a buyer in March 2023 futures $1800 call options for $13.75 at 2000X, odd trade	Macro	Stock Commentary
FinBERT	APPLE REACHES $50 MLN SETTLEMENT OVER DEFECTIVE MACBOOK KEYBOARDS - RTRS	Company | Product News	Legal | Regulation

Table 2: Sample text from each model with true label versus predicted label.

the highly technical training undertaken for FinBERT.

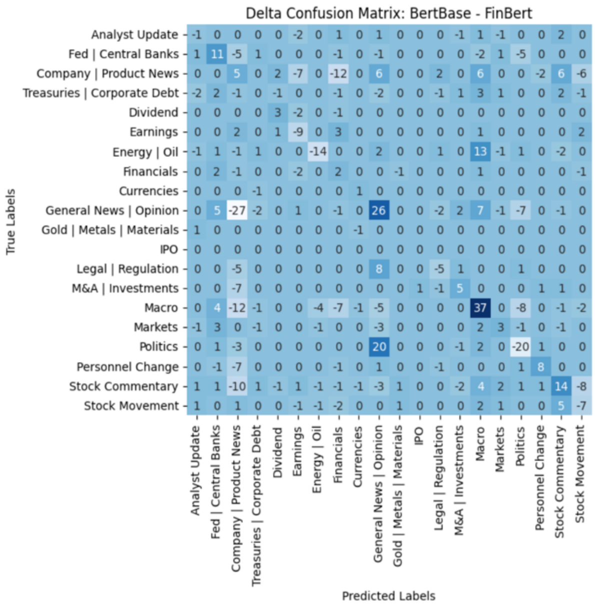

Figure 2: Confusion Matrix with the Delta of BERTbase (Fit) – FinBERT (Fit).

5 Conclusion & Future Experimentation

We endeavored to develop a robust and accurate model for classifying finance-related tweeter data. To accomplish our goal, we theorized that employing a FinBERT transformer would outperform a BERTbase transformer. Moreover, we believed that by implementing methods for further pre-training and fine-tuning, we would enhance the capabilities of these models.

Based on our research, we concluded that FinBert's classification ability is superior when classifying text without fine-tuning. However, after implementing further pre-training and fine-tuning, we discovered that BERTbase quickly surpassed the domain-specific model and had a better ability to handle certain categories, such as opinion and company news that FinBERT mislabeled.

Given these discoveries, we propose that future experimentation consider these key ideas. Firstly, when dealing with data such as tweets where opinions vary, consider using a multi-label classification structure to account for the overlap in label similarities. Secondly, when possible, work with larger data sets to give the fine-tuning of BERT models the information it needs to learn critical insights. Finally, consider using a two-model approach where BERTbase and FinBERT are responsible for predicting only the categories they excel at, then combining the results into one final classification.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Yang, Yi & UY, Mark & Huang, Allen. (2020). *FinBERT: A Pretrained Language Model for Financial Communications*, https://www.researchgate.net/publication/342198406_FinBERT_A_Pretrained_Language_Model_for_Financial_Communications

Lin, Y.-C., Chen, S.-A., Liu, J.-J., & Lin, C.-J. (2023, June 12). *Linear Classifier: An Often-Forgotten Baseline for Text Classification*, https://arxiv.org/abs/2306.07111

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019, May 14). *How to Fine-Tune BERT for Text Classification?*, https://arxiv.org/abs/1905.05583.

Shnarch, E., Gera, A., Halfon, A., Dankin, L., Choshen, L., Aharonov, R., & Slonim, N. (2022, March 20). *Cluster & Tune: Boost Cold Start Performance in Text Classification.* https://arxiv.org/abs/2203.10581.