Evaluation of FinBERT Performance on Multi-Class ESG Classification Task based on the MSCI Framework

She, Fong Wing

August 2022

Abstract

In this paper, the performance of FinBERT, the Large Language Model that adapts to the financial domain, on multi-class classification of ESG categories based on the MSCI framework is evaluated. ESG sentences that are used in this research are extracted from the CSR reports of companies in 11 different industries. The evaluation is based on comparisons with other 11 Deep Learning and machine learning algorithms, including Naïve Bayes, Logistic Regression, Linear SVM, Random Forest, MLP, CNN, LSTM, Bi-directional LSTM, GRU, and BERT. Through fine-tuning of all the mentioned models, FinBERT is found giving the best performance in the task, but still have its performance extremely close to BERT, and the per-class precisions of FinBERT are less stable than BERT.

Keywords: Natural Language Processing; Large Language Model; Environmental, Social, and Governance (ESG); MSCI Framework; Multi-Class Text Classification

1. Introduction

In recent decades, copious amount of research is done on the use of Deep Learning algorithms to solve problems in different industry domains. Specifically in the financial domain, Deep Learning algorithms are used to analyse financial texts, and researchers are developing ways to incorporate the use of it in financial decision makings. Early this year in 2022, a state-of-the-art Large Language Model (LLM) that adapts to the financial domain, FinBERT, has been developed by researchers, and has documented that it significantly outperforms the Loughran and McDonald dictionary and other machine learning algorithms (Huang, Wang, & Yang, 2022).

This LLM is customized based on the Google's BERT algorithm, and has been evaluated on sentiment classification, ESG classification, and forward-looking statements classification tasks. Among them, ESG classification was done using 4 classes: environmental (E), social (S), governance (G), and non-ESG (Huang, Wang, & Yang, 2022).

In this research, we will instead be performing ESG classification using 9 classes based on the MSCI framework, including classes: Climate Change (CC), Natural Capital (NC), Pollution & Waste (PW), Human Capital (HC), Product Liability (PL), Community Relations (CR), Corporate Governance (CG), Business Ethics & Values (BE), and non-ESG (N). Our dataset consists of 4000 manually labelled sentences, each class with 500 sentences respectively. And we aim to evaluate the performance of FinBERT on multi-class ESG classification task based on comparison with other Deep Learning and machine learning algorithms, including Naïve Bayes (NB), Logistic Regression (LR), Linear Support Vector Machine (SVM), Random Forest (RF), Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN),

Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM), Gated Recurrent Unit (GRU), and Bidirectional Encoder Representation from Transformer (BERT).

2. Deep Learning NLP Algorithms for Text Classification

Deep Learning as a subset of machine learning is a neural network that aims to mimic human brains, and it consists of multiple layers including an input layer, zero to multiple hidden layers, and an output layer. Deep Learning models are trained by optimizing the model parameters based on the minimization of prediction errors during training and validation. Prediction errors are obtained through comparing the predictions by the model and the actual labels of the dataset.

NB is a probabilistic classifier that is often used in text classification problem due to its scalability, simplicity, and effectiveness. However, its performance is known to be less reliable compare with other models (Kim, Rim, Yook, & Lim, 2002). LR uses the sigmoid function at the output layer which produces probabilities between zero and one. In text classification problem, it is also known as a great baseline supervised algorithm. Linear SVM on the other hand aims to maximize the margin between data classes using a linear classifier. It is widely known as one of the best text classification algorithms (S. Li, 2018). RF is known to be suitable for tasks that deal with high-dimensional data such as text classification. It consists of multiple decision trees which each of them is trained using some random subsets of features.

For Artificial Neural Networks (ANN), MLP consists of fully-connected layers and is the simplest ANN. CNN makes use of convolutional and pooling layers to extract higher-level features. Different sizes and numbers of kernels are used to represent N-

gram features in text which helps it to have promising performance in various NLP tasks. Recurrent Neural Network (RNN) on the other hand uses sequential data such as text and connects nodes as a directed graph along a word sequence. LSTM, Bi-LSTM, and GRU are types of RNN. LSTM has a feedback connection which remember outputs from previous nodes, allowing it to capture long-term dependencies between words in text sequences. Bi-LSTM on top of LSTM, can perform input flows also in a backward manner, capturing long-term dependencies in both forward and backward directions. GRU as a gating mechanism in RNN, the number of parameters is fewer than LSTM because it does not contain an output gate. All these models are common in performing text classification.

Besides, Large Language Models (LLM) such as the Google's BERT model is becoming a top trend in many NLP tasks, including sentiment analysis, question answering, text prediction, etc. It uses contextualized embeddings to represent words in text, which, words in different context are represented by different vectors. BERT was pre-trained using a massive dataset and the pre-trained model could be used in downstream tasks such as text classification by simply fine-tuning it. FinBERT on the other hand is a pre-trained model similar to BERT but has also adapted to financial domain through training with financial texts.

3. Methodology

We manually labelled sentences extracted from CSR reports of 55 companies, 5 companies from each industry, with a total of 11 different industries. The 5 companies are from different subgroups (Quintiles) within each industry to ensure variability in our dataset. A total of 4000 sentences is labelled, with 500 sentences per class. There

are in total of 9 classes, including Climate Change (CC), Natural Capital (NC), Pollution & Waste (PW), Human Capital (HC), Product Liability (PL), Community Relations (CR), Corporate Governance (CG), Business Ethics & Values (BE), and non-ESG (N). This dataset is then split randomly with 81% (3,240 sentences) for training, 9% (360 sentences) for validation, and 10% (400 sentences) for final model testing. All the fine-tuning is done using the validation set, and the same test set is used for all models. Testing is performed only at the very end of the tuning process of each model to evaluate their performance for comparison purpose.

For NB, LR, Linear SVM, and RF, text cleaning including converting words into lowercase, replacing symbols by spacing or deleting them, and stopwords removal is done before fitting the dataset into the models. For MLP, text pre-processing including tokenization, stemming, stopwords filtering, and extracting N-gram features is done before converting the sentences into one-hot vector for input to the model. For CNN, LSTM, Bi-LSTM, and GRU, same text pre-processing steps as MLP is used but without extraction of N-gram features, as well as sentences are converted into index vectors instead of one-hot vectors. All the 4 models include a word embedding layer after the input layer, and one or more fully-connected layers as the output layer. Finally, for BERT and FinBERT, their corresponding tokenizer is used for text pre-processing.

For all models, prediction label is determined by assigning each sentence to label with the highest likelihood predicted by the algorithms. Having the predicted labels from the model, we compare it with the actual labels to calculate the accuracy and macro-f1 score, and draw up the confusion matrix for performance evaluation purpose. Besides, fine-tuning of models are all based on the validation accuracy.

4. Results

Model	Test Accuracy (%)	Test Macro-f1 Score (%)
FinBERT	89.78	89.81
BERT	89.78	89.75
MLP	88.00	88.03
Linear SVM	84.89	84.83
Logistic Regression (LR)	84.00	84.09
CNN	83.78	83.96
Bi-LSTM	83.11	82.92
Naïve Bayes (NB)	82.89	82.99
Random Forest (RF)	81.33	81.50
LSTM	76.44	76.72
GRU	74.89	75.43

Table 1: Test accuracy and macro-f1 score of all 11 models

Table 1 above reports the 11 fine-tuned Deep Learning and machine learning algorithms' performances on the testing dataset (10% of the whole dataset, 400 sentences in total). It is found that FinBERT performs the best among all the other models. It has a test accuracy of 89.78% and a macro-f1 score of 89.81%. BERT has the second-best performance, with a test accuracy of 89.78% and a macro-f1 score of 89.75%, only a very small difference in macro-f1 score comparing with FinBERT.

Third is MLP, with a test accuracy of 88.00% and a macro-f1 score of 88.03%. These 3 models outperform all the other models by at least 3% test accuracy and macro-f1 score.

The rankings of the remaining models from highest to lowest (4th to 11th) based on the test accuracy are Linear SVM (test accuracy: 84.89%, macro-f1 score: 84.83%), LR (test accuracy: 84.00%, macro-f1 score: 84.09%), CNN (test accuracy: 83.78%, macro-f1 score: 83.96%), Bi-LSTM (test accuracy: 83.11%, macro-f1 score: 82.92%), NB (test accuracy: 82.89%, macro-f1 score: 82.99%), RF (test accuracy: 81.33%, macro-f1 score: 81.50%), LSTM (test accuracy: 76.44%, macro-f1 score: 76.72%), and GRU (test accuracy: 74.89%, macro-f1 score: 75.43%).

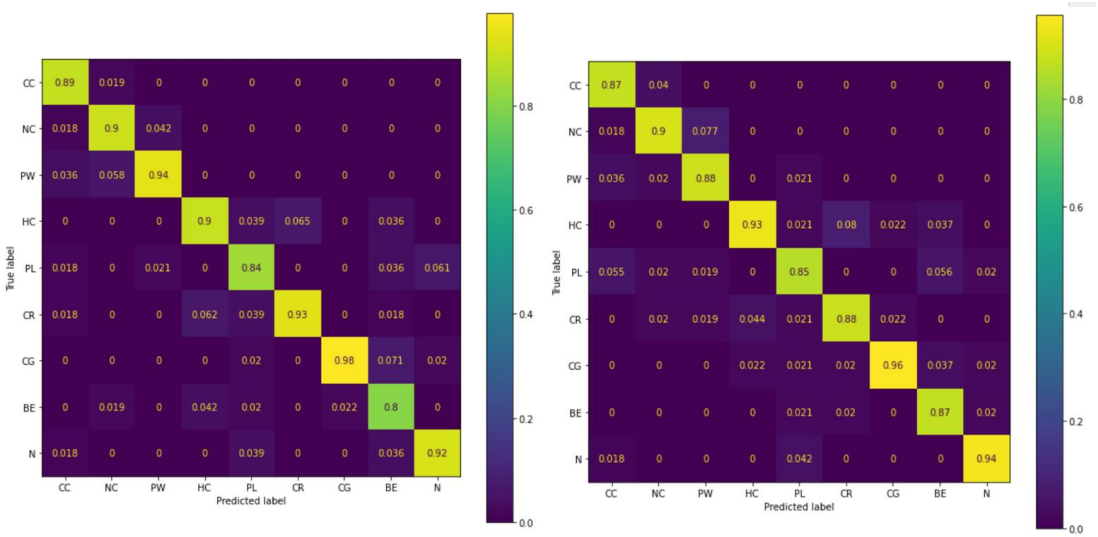

Figure 1: Left – Confusion matrix of FinBERT on testing dataset, Right – Confusion matrix of BERT on testing dataset

Above in Figure 1 are the confusion matrixes of the top 2 models, FinBERT and BERT, on testing dataset. On the left of Figure 1 illustrates that for FinBERT model, 89%, 90%, 94%, 90%, 84%, 93%, 98%, 80%, and 92% of the CC, NC, PW, HC, PL, CR, CG, BE, and N labelled sentences are predicted correctly by FinBERT respectively,

with an average macro-precision of 90%. On the right of Figure 1 illustrates that for BERT model, 87%, 90%, 88%, 93%, 85%, 88%, 96%, 87%, and 94% of the CC, NC, PW, HC, PL, CR, CG, BE, and N labelled sentences are predicted correctly by BERT respectively, also with an average macro-precision of 90%.

5. Discussions

From Section 4 Table 1, FinBERT outperforms all other tested Deep Learning and machine learning algorithms based on test accuracy and macro-f1 score. However, the performance of FinBERT and BERT is superbly close, with same test accuracy of 89.78% and a small 0.06% difference in test macro-f1 score. It does not demonstrate a strong advantage of FinBERT on classifying financial text compare with BERT. But the result still, has proven that FinBERT can achieve high performance similar to the BERT model, strongly outperforming other NLP algorithms.

From Section 4 Figure 1, the 2 normalized confusion matrixes have illustrated that the 2 LLM models, FinBERT and BERT, give the same average macro-precision of 90%. However, from the matrixes, it is also observed that FinBERT model tends to give more extreme per-class precision performance compare with BERT model. For instance, looking at the FinBERT model, the class that obtains the highest precision score is CG, with a precision of 98%, while the class that obtains the lowest precision score is BE, with a precision of 80%, that is with a huge 18% difference in precision. On the other hand, looking at the BERT model, the class that obtains the highest precision score is also CG, with a precision of 96%, and the class that obtains the lowest precision score is PL, with a precision of 85%, that is with a smaller 11% difference in

precision. It is therefore observed that FinBERT model gives a less stable precision performance in classifying sentences with different classes than the BERT model.

6. Conclusions

With the increasing demand of high-performance NLP models for solving problems and aid strategy planning in the financial domain, in this paper, the state-of-the-art LLM FinBERT is evaluated on a manually labelled dataset for multi-label ESG text classification task. Among all the 11 models experimented, including NB, LR, Linear SVM, RF, MLP, CNN, LSTM, Bi-LSTM, GRU, BERT, and FinBERT, FinBERT gives the best performance in test accuracy and macro-f1 score. However, the performance of FinBERT is superbly close to the BERT model, and it is also observed that FinBERT model tends to give less stable precision performance in classifying sentences with different classes than the BERT model, it does not demonstrate a strong advantage of FinBERT on classifying financial text compare with BERT. But the result still, has proven that FinBERT can achieve high performance similar to the BERT model, strongly outperforming other NLP algorithms.

Reference

Huang, A. H., Wang, H., & Yang, Y., “FinBERT—A Large Language Model Approach to Extracting Information from Financial Text”, 2022, pp.1

Kim S. B., Rim H. C., Yook D. S. and Lim H. S., “Effective Methods for Improving Naive Bayes Text Classifiers”, LNAI 2417, 2002, pp. 414-423

S. Li, “Multi-class text classification model comparison and selection,” Medium, 06-Dec-2018. [Online]. Available: https://towardsdatascience.com/multi-class-text-classification-model-comparison-and-selection-5eb066197568. [Accessed: 13-Aug-2022]