Assignment 6.1: Implementing and Evaluating a Large Language Model (LLM) for Text Classification

Patricia Enrique

Department of Engineering, University of San Diego

AAI-520: Natural Language Processing

Prof. Mokhtari

October 16, 2023

Data Loading and Model Selection

The dataset used for the text classification implementing large language models (LLMs) is compiled by Consumer Financial Protection Bureau (CFPB) (Tiwari, 2021) and is comprised of 162400 disputes between financial institutions and consumers. The disputes are categorized into five categories: credit reporting, debt collection, mortgages and loans, credit cards, or retail banking. The distribution of the dataset is highly imbalanced with 56% in the credit reporting class, and the remainder distributed between the remaining classes. The disputes are converted to a numerical scale and changed to lowercase before implementing the pre-trained Bidirectional Encoder Representations (BERT) model from Transformers. A sample of the disputes are shown in Table 1.

Table 1

Sample of disputes in the dataset

ID	Narrative	Product
16582	due corona pandemic work school taking online ...	Mortgages and loans
150888	account paid settlement erc account still show..	Debt collection
30723	extremely furious pray help get letter bureau ...	Credit reporting
129621	following account listed equifax credit report...	Credit reporting
115585	requested bank ca transfer bank account living...	Retail banking

The BERT model has been trained on a large corpus making it ideal for small, defined tasks. It is selected for its ability to obtain high accuracy with little fine-tuning and its ability to be used immediately for various tasks (Tunstall et al., 2020).

Model Evaluation

The pre-trained BERT model for sequence classification is loaded from the Transformers library and AdamW is set as the optimizer. To fine-tune the BERT model on the dataset, the loss and accuracy is measured for each epoch using the training and validation data (Dubey, 2020). The metrics do not significantly improve with the addition of epochs past two. Due to the time cost to run with higher

epochs, the model is limited to two. Once the fine-tuning is completed, the trained model is evaluated on the testing set using accuracy, precision, recall, and F1-score metrics. The final performance can be seen in Table 2. The model obtained an 88% performance on all the metrics with minimal fine-tuning. This indicates that 88% of the class predictions were correct and similarly, 88% of an actual class was identified correctly.

Table 2

Performance metrics for the trained model on the testing set

Model	Performance Metric			
	Accuracy	Precision	Recall	F1 Score
Testing Set	0.88	0.88	0.88	0.88

Challenges and Observations

A major disadvantage to the BERT pre-trained model is the model size and computational effort required to run it, and the training process is slow due to the number of weights that need to be updated. However, the BERT model is able to obtain high accuracy and can be fine-tuned and used immediately for various tasks (Tunstall et al., 2020).

In this analysis, the original dataset is too large to run in a reasonable amount of time. Instead, a sample of 100000 disputes are used. This combined with the uneven distribution of classes can be contributors to the lower performance metrics. Additionally, the BERT model has a maximum padding length which for this analysis is set to 128 tokens. After which the dispute is truncated. If important text classification information is contained after the 128 tokens, the model is not able to see it.

References

Dubey, A. (2020). BERT for sequence classification. *Kaggle*. https://www.kaggle.com/code/akshat0007/bert-for-sequence-classification

Tiwari, S. (2021). *Consumer Complaints Dataset for NLP* (Version 1.0) [Data set]. Kaggle. https://www.kaggle.com/datasets/shashwatwork/consume-complaints-dataset-fo-nlp

Tunstall, L., Von Werra, L., & Wolf, T. (2022). *Natural Language Processing with Transformers*. O'Reilly Media.