Cross Dataset Model Evaluation in NLP

Austin Chen

***Abstract*—In machine learning, it is often impossible to determine how well a model will perform on data that is different from the data the model was trained on. However, with the generalizability of natural language, it is assumed that a model that is sufficiently complex and receives sufficient training data can perform well in all cases. This project seeks to examine the generalizability of natural language by evaluating models on text datasets different from the training sets of the original models in a simple sentiment analysis task.**

I. Introduction

NATURAL language processing tasks enjoy a greater level of generalizability compared to other tasks in machine learning due to the generalizability of natural language itself. As data and labeling is generally uniform across NLP datasets, the question naturally arises of how well a model trained on one dataset may perform on another.

It is well known that models trained on very large text corpuses can generalize well across a variety of natural language tasks, and these models are often called Large Language Models (LLMs). LLMs take advantage of the sheer size of the dataset being trained on as well as a complex neural network with many layers to effectively learn the data in the text corpus [1] [2]. However, such data is not always available, and it can be difficult to determine whether a model trained on a smaller text corpus will be able to generalize in the same way across different styles and types of text.

Being able to generalize model performance on one dataset to performance on different datasets can alleviate issues where there is a lack of certain types of data in NLP but an abundance of other similar data. In those cases, it can be helpful to know how training on a different dataset may impact model performance.

II. Methodology

Considering the limited computational resources and access to datasets, a sentiment analysis task was determined to be more conducive towards gathering results on model performance across different datasets [4].

A. Datasets

Datasets were mainly gathered from two sources for this experiment: Amazon Review Data (2018) [5] and Financial Phrasebank [3]. Three datasets were extracted from the Amazon Reviews data, being the 5-core Video Games dataset, the 5-core Arts, Crafts and Sewing dataset, and the 5-core Office Products dataset. From the Financial Phrasebank dataset, the 50% agreement dataset was used.

After preprocessing the dataset, four more datasets were generated from the original datasets for a grand total of eight datasets. The first of these datasets was taking the Video Games review dataset and reversing each of the reviews in the dataset while maintaining the same label. The second dataset was created by taking the Video Games review dataset and shuffling the words in each review. The third dataset was created by taking the first in every 500 reviews in the dataset, resulting in a dataset that was 0.2% the size of the original dataset. The final dataset was obtained by taking alternating reviews from the Video Games review dataset and the Financial Phrasebank dataset until one dataset ran out of samples.

The eight datasets will be referred to as follows in the order of introduction:

- Video Games
- Arts, Crafts and Sewing
- Office Products
- Financial Phrasebank
- Video Games Reversed
- Video Games Shuffled
- Video Games Truncated
- Video Games & Financial

B. Preprocessing

Each of the models were preprocessed in a standardized way so that the results would remain generalizable. The steps were as follows:

1) The labels were standardized to a 2 for positive sentiment, 1 for neutral sentiment, and 0 for negative sentiment. For reveiw data, this applied to 4-5 star reviews, 3 star reviews, and 1-2 star reviews, respectively.
2) The words were tokenized into lowercase words with punctuation fully removed.
3) Stopwords were removed from all the samples in the dataset entirely.
4) All words were lemmatized to reduce the vocabulary size.

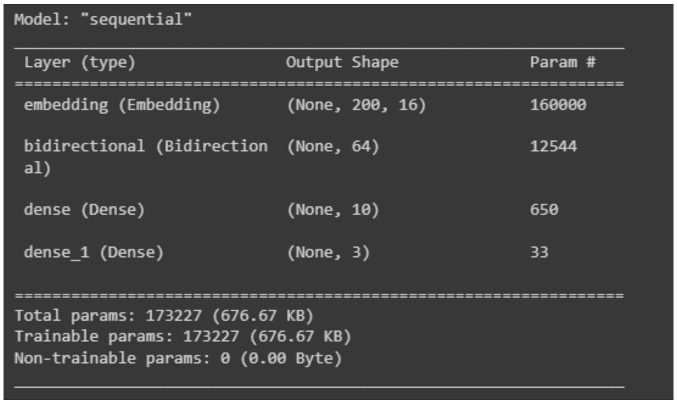
Before being fed into the model, each sample was tokenized into a numerical form with a maximum length of 200 words per sample and a vocabulary size of 10000.

C. The Model

The model was written in Tensorflow and consisted of the following layers:

1) An embedding layer of shape 200×16
2) A bidirectional LSTM layer with 64 outputs
3) A dense layer of size 10 with the ReLU activation function
4) A output layer of size 3 with the softmax activation function for classification.

The loss was taken with spare categorical cross entropy, and an Adam optimizer was used.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 200, 16)           160000

 bidirectional (Bidirection  (None, 64)                12544
 al)

 dense (Dense)               (None, 10)                650

 dense_1 (Dense)             (None, 3)                 33

=================================================================
Total params: 173227 (676.67 KB)
Trainable params: 173227 (676.67 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Fig. 1. Model summary as shown in Tensorflow

D. Training

The model was trained for 10 epochs maximum on a single dataset, with the with a patience value of 2 being applied with the validation set being the test set. As no other hyperparameters were being tuned, the model was not at risk of overfitting on the test set.

E. Testing

The model was evaluated first on its own test set for an initial accuracy value, and then it was evaluated on the entirety of each of the seven other datasets that it was not trained on. The resulting accuracies were tracked and later accumulated.

III. Results

The review data models were generally able to perform well on other types of review data, sometimes outperforming their accuracies on their own test sets. The reversed and shuffled datasets were also able to perform well on the other review datasets.

A. Metrics

The standardization of the preprocessing and training of the models allows for an opportunity to compare two datasets by comparing the accuracies they achieved on their own test sets and the other dataset. The resulting metric should satisfy two conditions:

- The metric should be inversely related to the accuracy that the model achieved on its own test set, because this value should be taken in comparison to the accuracy it achieves on the other dataset.
- The metric should be directly related to the accuracy that the model achieves on the other dataset, as the higher the accuracy achieved on the other dataset, the higher the similarity between datasets.

Following these conditions, a metric can be proposed called the ”Cross Evaluation Average” with the following formula:

$$CEA_{x,y} = \frac{acc_{x,y} + acc_{x,y}}{acc_{x,x} + accy, y} \tag{1}$$

where x and y are the datasets that are being compared and $acc_{a,b}$ is the accuracy of training a model on a and evaluating it on b.

B. Insights

We can attempt to generalize several trends based on the results of the data and the metric given above.

TABLE II
CEA: 0.9761

Trained on ↓ Tested on →	Video Games	Video Games Shuffled
Video Games	0.8656	0.8171
Video Games Shuffled	0.8589	0.8515

1) The Effect of Shuffling the Data: It is clear given the performance of the model trained on Video Games and the model trained on Video Games Shuffled that shuffling the dataset does not cause a large impact in terms of accuracy in sentiment analysis tasks in ths case.

TABLE III
CEA: 0.2983

Trained on ↓ Tested on →	Video Games	Financial Phrasebank
Video Games	0.8656	0.2586
Financial Phrasebank	0.2149	0.7215

2) Comparing Two Entirely Different Datasets: Here, each model performs extremely poorly on the other dataset, both achieving an accuracy of worse than 0.3333, or random guessing. It is clear that these datasets are very different from each other.

TABLE IV
CEA: 0.9223

Trained on ↓ Tested on →	Financial Phrasebank	Video Games & Financial
Financial Phrasebank	0.7215	0.5658
Video Games & Financial	0.7707	0.7276

3) Combing Two Datasets: Here, combining two datasets allows it to perform well on one of the datasets it was a combination of. In this case, the performance of Video Games & Financial was even better on Financial Phrasebank than the performance of Financial Phrasebank on Financial Phrasebank itself, perhaps due to the overlap of training sets. If we refer back to the main table containing all the accuracies, we can see that this combination is also able to perform well on the Video Games dataset.

C. Discussion

Many models were able to achieve higher accuracies on datasets they did not train on, and this was the most common with datasets being evaluated on the Arts, Crafts and Sewing dataset. One potential cause for this may be that the dataset

TABLE I
RESULTS OF EVALUATING THE MODELS ON DIFFERENT DATASETS

Trained on ↓ Tested on →	Video Games	Arts, Crafts and Sewing	Office Products	Financial Phrasebank	Video Games Reversed	Video Games Shuffled	Video Games Truncated	Video Games & Financial
Video Games	0.8656	0.9034	0.8849	0.2586	0.8194	0.8171	0.8994	0.5910
Arts, Crafts and Sewing	0.8348	0.9269	0.8994	0.3005	0.8087	0.8040	0.8390	0.5878
Office Products	0.8403	0.9208	0.9167	0.2780	0.7953	0.7935	0.8370	0.5764
Financial Phrasebank	0.2149	0.2297	0.2354	0.7215	0.2183	0.2173	0.2183	0.5658
Video Games Reversed	0.8254	0.8806	0.8564	0.2615	0.8706	0.8297	0.8260	0.5574
Video Games Shuffled	0.8589	0.9045	0.8860	0.2817	0.8627	0.8515	0.8531	0.5861
Video Games Truncated	0.7837	0.8846	0.8627	0.2810	0.7846	0.7867	0.7926	0.5600
Video Games & Financial	0.7095	0.7712	0.7315	0.7707	0.6899	0.6873	0.7082	0.7276

already contains very similar wordings, and the Arts, Crafts, and Sewing dataset may naturally be less noisy.

Many of the datasets were subject to overlapping training sets with other datasets. In particular, this was the case with Video Games, Financial Phrasebank, Video Games Truncated, and Video Games & Financial. Future work will need to examine the impact of this overlap on the accuracies that these models achieve on other datasets.

D. Future Work

The evaluation of models on different datasets provides a fast way for evaluation of the similarities between datasets.

1) Scarcity of Datasets: By comparing the performance of a model trained on a dataset in which the data can be easily obtained with the performance of a model trained on a dataset in which the data is scarce and difficult to obtain, one can assess the impact of introducing data from other datasets into the dataset that data is difficult to gather for.

2) Other NLP Tasks: This project only focused on the performance of LSTM models on sentiment analysis tasks. Future work can examine the performance of other models such as BERT or Word2Vec on other classification tasks or generative tasks.

IV. Conclusion

Cross Dataset Model Evaluation shows promise as a method of comparing datasets for their performance on various tasks. In this project, I focused on the performance of LSTM models on different datasets for sentiment analysis. Although this method is early in development, the results show promise for using smaller datasets to rival LLMs in performance on various specific tasks.

References

[1] Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models, 2023.

[2] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes, 2023.

[3] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65, 2014.

[4] Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32, February 2018.

[5] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics.