Generalizing Sentiment Analysis Models on User Data

Sathvika Anand and **Tonyaradzwa Chivandire**

Natural Language Processing
May 2, 2023

Abstract

We hope to explore whether trained sentiment analysis models are generalizable to other datasets, and how their performance may vary depending on the dataset. We found that some models performed well on other similarly themed data, and some even performed better than on their own data. The project exemplifies the impact that training sets can have on sentiment analysis models, and possible ethical implications this could have on more large scale projects.

1 Introduction

Our project is observing how different training data can affect the predicted sentiment of a text. For example, we want to explore if a model trained on movie review data can generalize to twitter data. To do this, we trained a sentiment analysis model separately on user data from 4 different sites: Amazon (product reviews), Twitter (Tweets about US Airlines), Reddit (Comments on political affairs in India), and RateItAll (Movie reviews). We then tested each model on the other datasets to see the differences in performance.

2 Setup

For our classifier, we used an RNN machine learning model with an embedding layer, LSTM layer, a dense layer and dropout mechanism between each layer to avoid overfitting. [1]

We had 4 models with this setup, one for each dataset (Amazon [2], Twitter[3], Reddit[4], RateItAll[5]). For each model, the training data consisted of 10K examples (80% training, 20%, validation) and test data consisted of between 1541 and 5000 examples, depending on how much data we had to split up. Each example in the dataset was labeled with a sentiment and all neutral reviews were eliminated to binarize our data to positve/negative sentiments. First, each model was trained on its own training dataset, and the training and validation accuracies were recorded, as seen in Figure 1 for the Amazon dataset. We also calculated the loss over each epoch of testing, as seen in Figure 2, again for the Amazon dataset. Then each model was tested on its own testing dataset, and then also tested on the other three test datasets. We calculated the accuracy to measure the performance of the each model by counting the number of correct sentiments and dividing it by the total number of sentiments made.

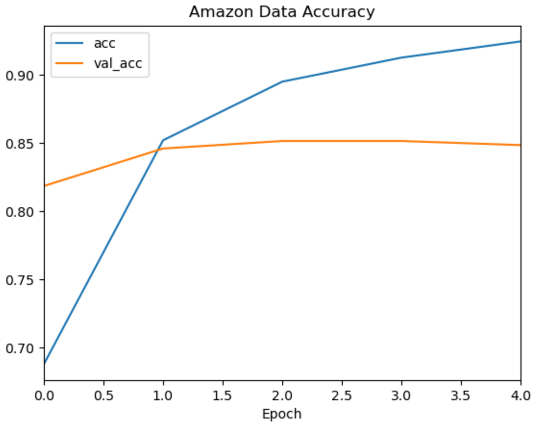

Figure 1: Example Accuracy plot for training a model on Amazon review data

3 Results

We see in Table 1 the accuracy calculations from running a trained model (rows) on a test dataset (columns).

[1] Techvidvan article: https://tinyurl.com/2p8r6des

[2] Amazon Reviews Sentiment Analysis Dataset: https://tinyurl.com/bpapkxxx

[3] Twitter US Airline Sentiment: https://tinyurl.com/3ks7xw6c

[4] Reddit Sentiment Analysis Dataset: https://tinyurl.com/2a7hr4tu

[5] movies.data from Assignment 7

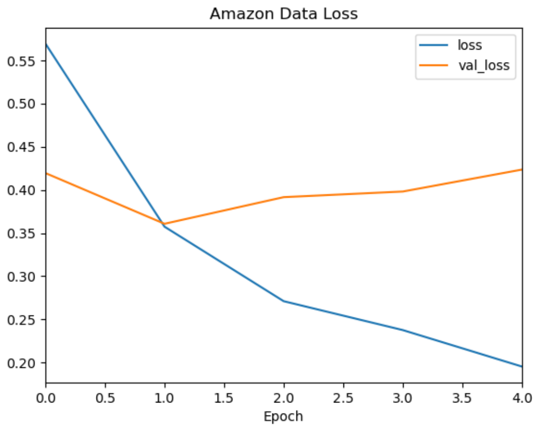

Figure 2: Example loss plot for training a model on Amazon review data

Train Data/Test Data	Amazon	Twitter	Reddit	RateItAll
Amazon	0.853	0.233	0.608	0.880
Twitter	0.331	0.911	0.507	0.385
Reddit	0.520	0.618	0.594	0.664
RateItAll	0.755	0.388	0.670	0.939

Table 1: Accuracies of models on differrent datasets

4 Conclusions

From the results above, we see that most models performed the best on datasets similar to those that they were trained on. However, we can see some interesting results with the Reddit-trained model, which actually did better on the movie reviews and twitter datasets than on its own set. We think this is because the Reddit dataset contained a wide variety of comments, in very colloquial language, which made it harder for it to predict sentiments. However, the Twitter and Movie review datasets were also mostly written in colloquial language but were more targeted topics, which might have made it easier for the model to predict a sentiment. We also noticed that the Amazon-trained model did quite well in predicting sentiments of movie reviews, probably because both datasets were consumer reviews, and likely contained similar opinion words.

5 Ethics

From our specific project, we can see the importance of how models are trained, and the effects of running models on data they were not trained on. It is clear that the data that a model is trained on is incredibly important for the results that it can generate. This may not be as impactful given our datasets in this project, but for other applications we need to keep in mind that data can carry inherent biases that are then reproduced when we test on a more general corpus. More generally, sentiment analysis trains the computer to recognize both positive and negative sentiment. AI chatbots that rely on sentiment analysis models could potentially produce or suggest toxic content based on the negative sentiment data they are trained on. Additionally, a faulty sentiment analysis model could falsely recognize a certain sentiment and make decisions based on it, for example labeling a certain company as "bad" based on an abundance of negative reviews.

6 Appendix

6.1 References

TechVidvan Tutorials Sentiment Analysis using Python
https://techvidvan.com/tutorials/python-sentiment-analysis/

Amazon Reviews Sentiment Analysis Dataset
https://tinyurl.com/bpapkxxx

Movie Reviews Sentiment Analysis Dataset from www.rateitall.com, as used in Assignment 7.

Reddit Sentiment Analysis Dataset
https://tinyurl.com/2a7hr4tu

Twitter US Airline Sentiment
https://tinyurl.com/3ks7xw6c

6.2 Contributions

Both of us contributed to the project equally - we were each in charge of preprocessing 2 out of the 4 datasets that we had, and then Tonya made the individual datasets while Sathvika rearranged the code to run all models in one go. Then we both worked together to calculate accuracy scores.

6.3 Hours Worked

See Table 2 below.

Date	Hours	Notes	Name
4/16/23	0.5	working on project presentation	both
4/18/23	1	worked on project writeup	both
4/22/23	2	getting tensorflow setup + initial sentiment analysis model working	both
4/24/23	2	worked on amazon review data	Sathvika
	1	debugging tensor flow installation	Tonya
4/28/23	1	syncing up, cleaning code up, deciding next steps	both
4/30/23	3	created individual datasets, combined code to run all four models at once, calculated accuracy, ran experiments	both
5/1/23	2	worked on paper draft, ran code to get results, made presentation	both
5/2/23	1	practiced presentation, finished paper	both

Table 2: Hours worked: Total = 14.5

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Figure 3: ...