Taylor King, Michael Moore, Ashlyn Woodward, Caleb Yost

Dr. Sayed Khushal Shah

CSCE 5290.001 NLP

Project Proposal

Generating and Evaluating Synthetic Reviews

1 Increment Two

1.1 Introduction

For increment two, our group focused on completing the models for our project and documenting our results. As described before, our project consisted of two main models, the first to generate synthetic reviews, and the second to classify reviews as real or fake. We used the Amazon reviews dataset as training for the first model in order to create the new reviews. We then fed the new reviews into our classifier model to see if it could detect the difference between the reviews. A diagram further showing the flow of our project can be found below in Figure 1.

1.2 Background

The background for this project stems from our observation of demand for this service within the global market. Labeled as the, "152 billion dollar problem," fake reviews take up an estimated 4% of all online reviews. Although many consider populating items with fake reviews unethical, the market for these reviews is undeniably large. Often these reviews are copy and pasted over many profiles and easily detectable. Our motivation is to make a model to generate unique reviews based on the product's information. A similar project to this can be found at this site [2]. However, with this large market the use of these reviews can be unethical. Fake reviews on a product can be the deciding factor if a consumer will purchase the item. For this reason our group will also design a model to help identify the fake reviews. A similar project to this portion can be found here [6]. This project will aim to help combat the epidemic of fake reviews that we have been facing since the boom of the online market.

1.3 The Models

Our project contained two models: a generative model and a classification model. The generative model was made using the aitextgen package in python; a package that creates AI models based of the GPT-2 architecture. This model was trained for 5000 steps with a learning rate of 0.001 and a batch size of 1. Overall, the results of this model exceeded expectations and generated seemly realistic reviews. The second classification model was based on the BERT model with a linear classification head added on the top of it. This model was implemented using the huggingface package in python in conjuction with pytorch. This model was trained for 4 epochs with a learning rate of 0.00002 and a batch size of 64. This model performed better than expected, with an Matthews Correlation Coefficient (MCC) of 96% on a held-out test set.

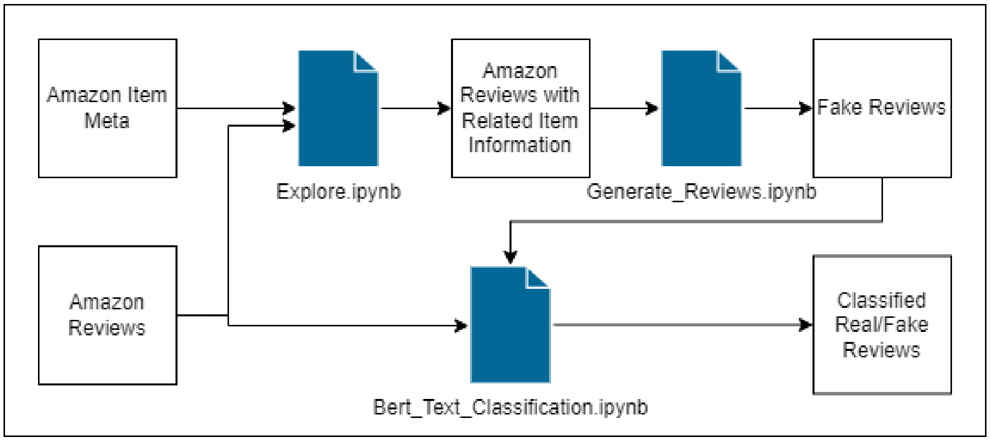

Amazon Item Meta
Explore.ipynb
Amazon Reviews with Related Item Information
Generate_Reviews.ipynb
Fake Reviews
Amazon Reviews
Bert_Text_Classification.ipynb
Classified Real/Fake Reviews

Figure 1: Project Architecture and Workflow

1.4 Dataset

As stated in our iteration 1 paper, this project takes use of the Amazon Review dataset used in these publications [3] [4]. Originally we planned to use the majority of the dataset to accomplish our goal. After working with the dataset however, we discovered that it was too much data for us to train our models in a reasonable amount of time. This lead us to shrink our dataset, only using the Musical Instruments section of the dataset, which is comprised of around 1.5 million reviews. The dataset structures is listed down below in table 1.

1.5 Analysis of data

This project did not require much analysis of our dataset due to the nature of it. That being said, we did explore the rating's distribution in the Amazon dataset, as shown in figure 2. We noticed that the reviews were heavily skewed towards the positive end (reviews being 4 stars and up). This discovery lead us to hypothesis that our generated reviews would hold a positive sentiment.

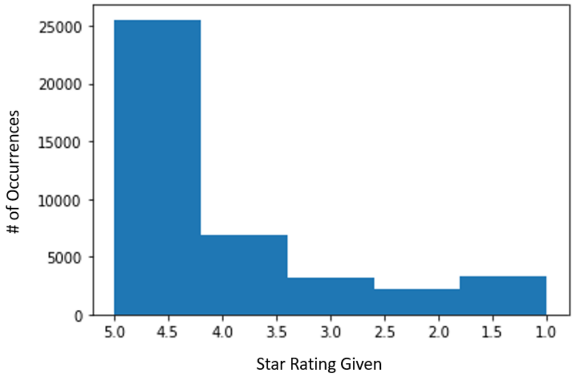


Figure 2: Distribution of Stars from the Amazon Ratings

1.6 Implementation

Before we could feed the Amazon review data into our generative model, we had to modify it slightly. The first thing we did was combine the two portions of the Amazon review data into a single dataset that contained the item information and the review of the item. This allowed us to generate an input into our model that contained the review and item information. Once compiled, we feed the full raw data into the model. We then discovered that the full data was too large for a model to be trained quickly and within RAM constraints. To mitigate this issue, we limited our training instances to 500,000 instances.

Feature Name	**Origin**	**Description**	**Example**
asin	Reviews & Metadata	ID of the product	0000031852
title	Metadata	Name of the product	TUTU Botiquecutie TM Girls Ballet Zebra Hot Pink
price	Metadata	Price of the product	$17.99
imurl	Metadata	Product image url	`https://m.media-amazon.com/images/I/718HTZfwz6L._AC_SX679_.jpg`
related	Metadata	Products related to the product	Also bought: {“B00JHONN1S”} Also viewed: {“B002BZX8Z6”}
salesRank	Metadata	Sales rank information for product	{”Toys & Games”: 211836}
brand	Metadata	Product Brand	Coxlures
categories	Metadata	Product sales categories	[[”Sports & Outdoors”, ”Other Sports”, ”Dance”]]
reviewerID	Reviews	ID of the reviewer	A2SUAM1J3GNN3B
reviewerName	Reviews	Name of reviewer	J. McDonald
helpful	Reviews	Helpfulness of review	2, 3
reviewText	Reviews	Review Text	”I bought this for my husband who plays the piano. He is having a wonderful time playing these old hymns. The music is at times hard to read because we think the book was published for singing from more than playing from. Great purchase though!”
overall	Reviews	Rating of review	5.0
summary	Reviews	Summary of review	Heavenly Highway Hymns
unixReviewTime	Reviews	Time of review (unix time)	1252800000
reviewTime	Reviews	Time of review (raw)	09 13, 2009

Table 1: Amazon Dataset Features to be used in our Analysis

Once we trained our generative model, we collected 1,000 instances of fake reviews. Examples of these were: ”Not a lot to say about.”, ”ok”, ”I bought this strap a few years ago. I love Ernie Ball straps. The straps are comfortable and they will last for years as far as possible. There are no issues with this strap. I think I have to return it back a little, but it is my second one. I would recommend it for any other guitar player. I own a Gibson Les Paul and it fits any Gibson Les Paul Les Paul Les Pauls.”, and ”It’s not a piece of any fancy design.”. These 1000 samples were then collated with real reviews which were then feed into the second classification model that was built following a guideline form OpGenious[5].

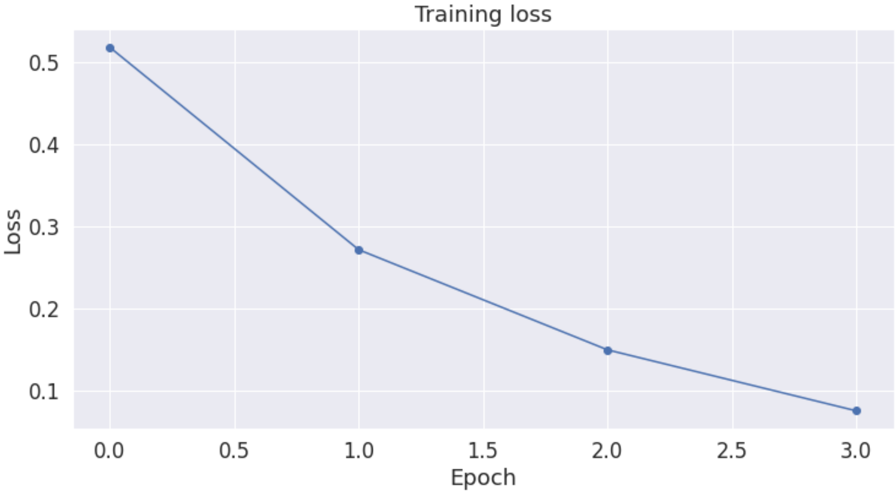
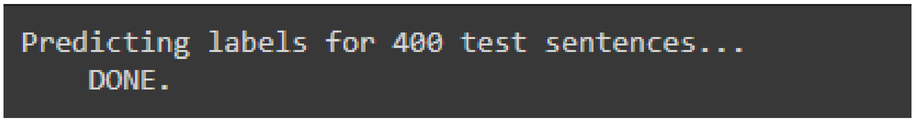


Figure 3: Training Loss of Text Classification Model

1.7 Results

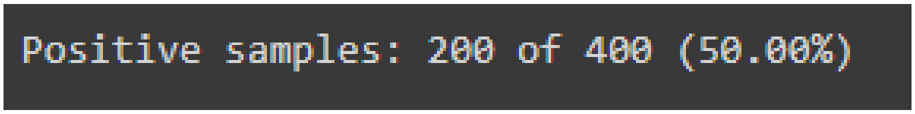
When training the classification model, we used 1600 total reviews, 800 real and 800 fake . Then for testing we used 400 reviews, 200 real and 200 fake. The training loss can be seen in Figure 3. When making the final predictions, the model outputs the text shown in Figure 4 to indicate it has made its predictions, it then outputs the text shown in Figure 5 to show how many reviews were classified as ’positive,’ or fake.

We used the Matthews Correlation Coefficient (MCC) to measure the quality of our model’s results. This statistical rate only produces a high score if the true positive, true negative,



```
Predicting labels for 400 test sentences...
    DONE.
```

Figure 4: Output to Show Model is Predicting Labels



```
Positive samples: 200 of 400 (50.00%)
```

Figure 5: Output Showing the Number of Reviews Classified as Fake

false positive, and false negative scores get good results, which makes it a reliable measure for a model[1]. The MCC for our classification model was 96.6%, which is much higher than we were expecting it to be.

1.8 Project Management

In terms of contribution, all four members of our team contributed equally, 25 percent. Clay and Taylor created the model to generate the synthetic reviews. Ashlyn and Caleb created the model to classify reviews as real or fake. We each then collaborated on the other models to ensure everything was running correctly and to integrate them for the final results. We originally wanted to feed more information from the first model to the second, such as the review rating or title, however we ran out of time for that portion. With more time we could have also created a larger dataset of fake reviews to feed into the classifier model to see if it could uphold its high accuracy. All the work can be found in our Github repository.

References

[1] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 2020.

[2] Pranav Harathi. Generating fake product reviews, Dec 2018.

[3] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517, 2016.

[4] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

[5] Aryanshu Verma. Application of bert : Binary text classification. *OP Genius*, Oct 2020.

[6] Kessie Zhang. How to detect fake online reviews using machine learning.