Taylor King, Michael Moore, Ashlyn Woodward, Caleb Yost

Dr. Sayed Khushal Shah

CSCE 5290.001 NLP

Project Proposal

Generating and Evaluating Synthetic Reviews

1 Project Summary

Over the last 30 years the world has seen a major shift from in person shopping in local market places to online shopping. While this change is convenient for most individuals it has come with some unforeseen risks to consumers. One such risk that is given to the consumers is finding reputable producers from which they can purchase their goods. The most common metric that consumers use is the reviews of a seller from previous customers that the seller has interacted with and also the quality of the product that is being sold. With such large trust in this metric there has been an epidemic in the online shopping community in the form of fake reviews using trained neural networks to gain consumer trust.

To summarize this project our group will be making two models. The first model will be trained on genuine product reviews found from amazon. This model will out pout fake reviews that should be indistinguishable form the real reviews that were used for training. Then we will take these fake reviews generated by our first model and train a second model to identify fake reviews. This will be done to help combat the problem that diminishes consumer trust in online shopping.

The motivation for this project stems is driven from our observation of demand for the product within the global market. “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. However, with this large market the use of these reviews can be unethically. 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. This project will aim to help combat the epidemic of fake reviews that we have been facing since the boom of the online market.

2 Goals and Objectives

The significance of this project comes from the amount of money that can be made from the application of both models. The market for fake reviews is estimated to influence $3.8 trillion of global e-commerce spend in 2021, according to Jonathan Marciano from theprint.com [3]. A model that creates non generic and unique reviews is extremely valuable for the arguably unethical market. In order to combat this, models which can tell the difference between real and synthetic reviews are also exceptionally beneficial. Especially since most companies which buy fake reviews use them for low quality, previously one-star products in order to boost those sales [2]. This project will help solidify the lessons taught in class by practically working with a text-based dataset and a neural network.

The objective of our project is to generate fake reviews for a product and to build a model to identify whether a given review is fake or not. We will start this by making a machine learning model that generates fake reviews. to accomplish this, we will be using the Amazon review dataset used in these publications [1] [4]. All features and an example for each can be found in Table 1. Once we create fake reviews, we will be able to move to the next part of our project of identifying them with a machine learning model. This part of the project will be done via supervised learning, labeling the reviews from the amazon dataset as true and our synthetic dataset as false. By the end of our project, we ultimately hope to have a model that can recognize the synthetic reviews with an accuracy of at least eighty percent and a model that can generate realistic-looking reviews. These goals conflict with each other, so for the purposes of this class we will keep our focus mainly on the identification of the reviews.

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

References

[1] 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.

[2] Sherry He, Brett Hollenbeck, and Davide Proserpio. The market for fake reviews. *Available at SSRN 3664992*, 2021.

[3] Jonathan Marciano, Apoorva Mandhani, Shubhangi Misra, and Shobhaa De. Almost 4% of all online reviews are fake. their impact is costing us $152 billion, Aug 2021.

[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.