Fake Google Reviews in Alaska Using Data Mining

Project Milestone 1

Nick Sietsema
sietsemn@colorad.edu

Rebecca Toland
reto6656@colorado.edu

Amber Tin
amti9774@colorado.edu

Problem Statement / Motivation

Reviews are heavily used for purchasing decisions and those that are fake can be detrimental for customers and businesses. According to a 2016 study by the Pew Research Center, 82% of U.S. adults say they at least sometimes read online customer reviews before purchasing items for the first time, and 40% of that group say they almost always or always do so. Businesses can purchase fake reviews to improve their ratings, search results, and sales. Although the FTC is cracking down on review fraud, artificial intelligence tools threaten to compound the problem by cranking out far more fake reviews that appear highly authentic. Therefore, efforts to eliminate fake reviews is imperative in order to protect customers and support honest businesses.

The purpose of our project is to detect and classify online reviews into real or fake through feature engineering and rule based classification. We will validate our model using a test dataset that has real and fake review classes.

Literature Survey

Several research studies have been conducted on the detection of fake Google reviews using data mining techniques.

In the study conducted by Pendyala, A (2019) an extensive investigation was undertaken to address similar fake reviews, particularly through the Google platform. The research used many techniques including Naive Bayes, Linear SVC, Support Vector Machine (SVM), Random Forest, and Decision Trees algorithm. To enhance the accuracy even further, features like the sentiment of the review, verified purchases, ratings, emoji count, product category with the overall score were used. With implementing these methods, this study was able to get a success rate of over 79%.

In the research conducted by Hossain, F (2019), various algorithms were carried out to detect fake reviews including Support Vector Machines (SVM) with stochastic gradient descent (SGD) learning, Multinomial Naive Bayes, and Multi-Layer Perceptron with one hidden layer (MLP1) and two hidden layers (MLP2).

Salminen, et al (2022), found that using large language models (LLMs) to detect fake reviews had a near perfect success rate in comparison to human reviewers, which had a success rate of around 50%.

Ott, et al (2011) studied a different perspective than the other mentioned in this section. They primarily focused on using genre identification, psycholinguistic deception detection, and text categorization. With these approaches, they introduce techniques including Naive Bayes and Support Vector Machine classifiers. The success rate of these combined features led to an accuracy of nearly 90%.

Proposed Work

Data cleaning and preprocessing work will include removing noisy data, such as duplicate reviews that have the same user id, business id,

rating, text, and timestamp. This will indicate erroneous reviews. Duplications such as the same user id and business id, but a different timestamp will be retained for our analysis. We will also remove records with empty cells/ratings. Irrelevant data such as business replies will be removed as well. Rating will be converted to a numeric format and ratings aggregates will be added.

Will will use reviewer, review, and business-centric feature engineering to develop rules and classify the data. For reviewer-centric features, we will determine the average rating and maximum number of reviews in one day (e.g. users who post multiple reviews in one day will be flagged). Review-centric features will include word count, punctuation count, and key words. Business-centric features will include the maximum reviews received in one day, the average number of words in reviews, and average business rating. We will review our data to determine averages and set thresholds for these areas. The rule set will be prioritized to create a decision list. The class with the highest priority will be assigned as the final class and the sequential covering algorithm will be used for classification.

We will all be involved in developing the rule set, processing the training data, and testing the algorithm. Amber will work on the data cleaning and preprocessing. Rebecca and Nick will be in charge of the visualizations.

Dataset

We will use the google local review dataset for 2021 in Alaska to find fake reviews and flag them.

Dataset URL: https://datarepo.eng.ucsd.edu/mcauley_group/gdrive/googlelocal/

- 12,774 businesses in dataset
- 1.05 million reviews

Metadata:

- user_id - ID of the reviewer
- name - name of the reviewer
- time - time of the review (unix time)
- rating - rating of the business
- text - text of the review
- pics - pictures of the review
- resp - business response to the review including unix time and text of the response
- gmap_id - ID of the business

A subset of Amazon reviews (electronics category) will be used as the test dataset:
URL:
https://huggingface.co/datasets/amazon_us_reviews/tree/refs%2Fconvert%2Fparquet/Electronics_v1_00

Evaluation Methods

We will evaluate our project in terms of how well it spots fake reviews, using an Amazon reviews dataset (electronics category). The reviews include verified and unverified purchases from Amazon. Unverified reviews are those with no proof that the customer purchased the product on Amazon or elsewhere. A 2019 investigation by the U.K. consumer advocacy group Which? found that hundreds of unverified product reviews were being posted on Amazon product pages in a single day. Which? searched for 14 tech products and 71% of the products on the first page of the search results had 5-star reviews, however 90% of those reviews were unverified. More than 10,000 reviews from

unverified purchasers were discovered for just 24 items in a couple of hours. The organization described this as “an easy-to-find red flag,” indicating these are likely fake reviews.

We will test our algorithm’s ability to detect fake reviews based on how it classifies the unverified versus verified purchases from the Amazon dataset. For our evaluation, we will use metrics such as accuracy, precision, and recall on the Amazon test dataset. We will conduct a literature survey on previous work and highlight the key differences in our proposed work compared to previous work. Finally we will visualize our results looking at the outliers using a regression model.

Tools

Python 3 and NumPy will be used for analyzing and classifying the data. We will use Python Pandas for the data cleaning and preprocessing. Matplotlib will be used for data visualization during the knowledge discovery process and to present our findings. Collaboration will be through Github.

Milestones

1. By July 24, 2023 we will have data cleaning, preprocessing, feature extraction and engineering, and model development. We will also gain better knowledge on what our classifications will be within our model.
2. By July 31, 2023 continue feature extraction and engineering, evaluate the performance of the model on the testing set using our evaluation metrics like accuracy, precision, and recall.
3. By August 7, 2023 we will fine-tune and refine our model based on the findings evaluating the strengths and weaknesses. We will also fine-tune the models parameters and feature selection to improve performance. We will start working on our project presentation video. The project code and descriptions will start to be loaded into Github if not already loaded.
4. By August 14, 2023 the model will be finished and ready to submit.

1. Milestones Completed

We encountered a major hurdle in that our initial selected dataset was far too large to be loaded into main memory on any of our local machines or coding.csel.io. Initially, we discovered that we could stream the file using the chunksize argument, but even with this method simple aggregation operations still took minutes and permanent alterations to the dataset took upwards of ten minutes. The size of the dataset also precluded effective version control. We chased down further leads including tools for processing and versioning large datasets such DVC and Pachyderm, etc., but given the timeline of the project, we found these infeasible.

Given these challenges, we made the decision to work with a smaller dataset in order to press forward with the project. We switched from the Colorado dataset to the Alaska dataset, which was still too large to be loaded into main memory on any of our available resources. Fortunately we discovered that pandas can open compressed files, and by doing that in tandem with streaming and data reduction we were able to get the dataset down to a workable size. It was still too big for coding.csel.io, but our local machines can handle it.

Through strategic attribute selection, Nick successfully reduced the size of the Alaska JSON file from 89.8 MB to 68.5 MB. The revised dataset included user_id, name, time, rating, text, and location. Our team continued to work on the JSON file until July 24th which we

encountered several challenges including using that big of a dataset within our platforms.
This week, a pivotal step was taken as our team successfully converted the JSON file to a CSV file therefore streamlining its compatibility with Python for more efficient data processing. With the time crunch due to the conversion, data cleaning processes ensued, including the conversion of the column names to enhance the dataset's usability, dropping null cells, removing complete duplications, and converting the unix time to a readable form.

The data transformation process involved creating new attributes for counting the number of words in each text review and categorizing it based on ranges for analysis. Additionally, an attribute was added for the average number of reviews per day for each reviewer.

Two of the indicators of fake reviews are text entries with less than five words and reviewers that post more than five reviews in a day. The first indicator has been solved, and the second one is in progress; we are able to group the data by date, and are working on how to extract the names from each group that have more than five occurrences. This should be figured out by July 24, 2023, which will allow us to move forward with our model.

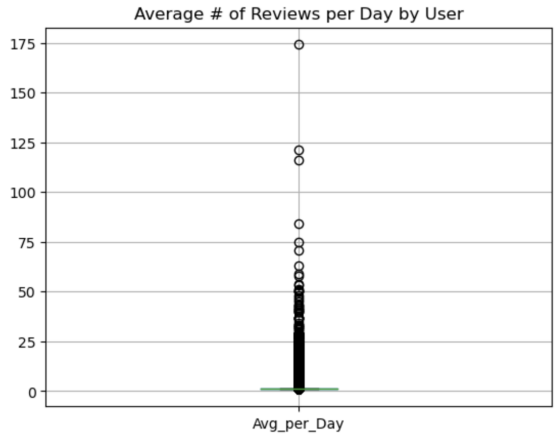
2. Milestones Todo

Due to the unforeseen delays encountered during the conversion process of our dataset, our team now faces a considerable workload ahead. The crucial tasks needed to be completed include more data cleaning, preprocessing, and feature extraction. To foster collaboration, our team has scheduled regular Zoom meetings, and continue to communicate through Slack, while the latest code updates and revisions will be managed through GitHub.

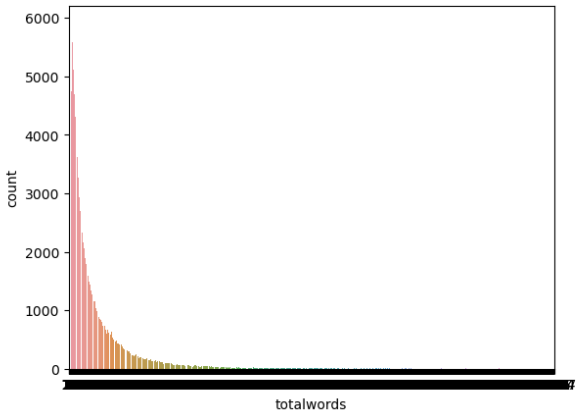
3. Results

The team started the data mining efforts, focusing on analyzing reviews per day per user. The examination led to the identification of an account that raised suspicion due to having 175 reviews in a single day – an observation suggestive of potential fake reviews.

User/day Reviews Example



As a part of our classification analysis, we wanted to investigate word count patterns. The graph below is a count plot graph, which will aid us in establishing optimal parameters for the word count analysis.



Word Count/Review Count Graph

Subsequently, a correlation graph was generated to look at the interrelationships among our different attributes. Surprisingly, the findings lacked any strong correlations between the attributes. Instead negative correlations were found, such as “total words” and “rating”, “total words” and “time”, as well as “total words” and “user IDs”. This observation will have to be further investigated.

Attribute Correlation Heatmap

References

F, Hossian. 2019. Fake Review Detection Using Data Mining. MSU Graduate Thesis. https://bearworks.missouristate.edu/cgi/viewcontent.cgi?article=4454&context=theses

M. Ott, Y. Choi, C. Cardie, J. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 21st international conference on World Wide Web,* pages 309-319, Association for Computational Linguistics,2011.https://aclanthology.org/P11-1032.pdf

A, Pendyala. 2019. Fake consumer review detection (Master’s thesis). California State University, Sacramento, Department of Computer Science.https://scholarworks.calstate.edu/downloads/xg94hv34z

J. Salminen, C. Kandpal, A. Kamel, S. Jung, B. Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services,* Science Direct. https://aclanthology.org/P11-1032.pdf

A, Smith, M. Anderson. 2016. Pew Research Center: Online Reviews. https://www.pewresearch.org/internet/2016/12/19/online-reviews/

2019. BBC: Amazon ‘Flooded by Fake Five-Star Reviews’ - Which Report. https://www.bbc.com/news/business-47941181