1866 AUB
American University of Beirut
الجامعة الأميركية في بيروت

OSB
SULIMAN S. OLAYAN
SCHOOL OF BUSINESS

MSBA 315 – Spring 2023

Project Presented to

Dr. Wael Khreich

Title: Predicting Amazon’s Fake Books’ Reviews

Presented by

Anthony Aramouny

Charbel Daccache

Jana Hasan Kassem

Roula Irani

Abstract

Customers who rely on reviews to make informed purchases are facing an increasing problem in light of the growth of fake reviews in today's internet marketplace.

In this project, we have provided a framework to detect Amazon's fake books' review. We have used content from Kaggle that wasn't labeled and developed a supervised machine learning model to detect the fake reviews. We applied CountVectorizer to label our data. Then we tried multiple models like Logistic Regression, KNN, Random Forest Classifier, Light gbm, and MLP Classifier to check the accuracy. Results from experiments with our proposed framework in a binary- classification are promising both with classifier accuracy, f-measure metrics, confusion matrix metrics. These results indicate that our proposed framework provides a feasible solution to detect Amazon fake books' review.

Introduction

The growing popularity of online shopping has experienced an important growth nowadays, with e-commerce companies such as Amazon, Alibaba, and eBay providing customers with a convenient and cost-effective possibilities for buying items.

Amazon's huge online marketplace has changed the publishing business by enabling writers to self-publish and reach readers all over the world. However, as self-publishing has grown in popularity, so too has the platform become a magnet for fake reviews. These evaluations, which are typically the product of paid review firms, artificially inflate the book's average rating and mislead consumers.

In this report, we aim to address the problem of fake book reviews on Amazon and predict the likelihood of a review being fake. What factors contribute to fake book reviews on Amazon, and can we develop a model to accurately predict the likelihood of a review being fake?

To answer this question, we will analyze a dataset of Amazon book reviews from Kaggle and use machine learning techniques to identify patterns and characteristics of fake reviews.

Ultimately, our goal is to provide valuable insights that can help Amazon improve its review system and protect consumers from being misled by fake reviews.

Literature Review

Related work to label our dataset, we found a research paper on 'ResearchGate' named 'Identifying fake Amazon reviews as learning from crowds'. Tommaso Fornaciari and Massimo Poesio used the Learning from Crowds algorithm as a reliable tool for labeling the reviews, so that effective models can be trained in order to classify them as truthful or not. The paper proposes a number of features to distinguish fake reviews from genuine ones, including the text of the review, the reviewer's history, and the temporal and spatial patterns of the reviews. In our dataset we couldn't find the same features so we used the review length and average rating. The paper uses a small labeled dataset of fake reviews to train a classifier, which is then used to identify fake reviews in a larger set of unlabeled reviews. We used the same labeled dataset and named it as a corpus. Then, we selected only two common features from our dataset and theirs that were rating, and review/text. As for the verified purchase feature we dropped this column since it doesn't exist in our dataset. After several trials to find a verified purchase in order to link it to our dataset, we went with the option of training the model without having this feature. The impressive thing is that while training the logistic regression on their data we only had an accuracy of 65%, and after implementing the LazyClassifier we found that the light gbm had the highest accuracy while taking into consideration the time taken for execution (5.83 s).

More related work to detect fake reviews has primary focused on review-centric characteristics using logistic regression. From 'International Research Journal of Engineering and Technology', Kaushik Daiv, Mrunal Lachake, Prathamesh Jagtap, Srishti Dhariwal, and Prof. Vitthal Gutte proposed in their paper called 'An Approach to Detect Fake Reviews based on Logistic Regression using Review-Centric Features' that labeling fake reviews with "verified purchase" works well. This study provides a logistic regression framework for fake review identification research, they also suggested two feature extraction methods, Tf-idf and CountVectorizer, and found that logistic regression using CountVectorizer on the dataset achieved 82% accuracy and 81% with Tf-idf. In our dataset we used from this research the CountVectorizer since it's the optimal feature extraction method comparing to tf-idf. We split our Review/Texts feature into vectors with max number of features of 1400.

We found in 'Journal of Business Research' that 'SuleO¨ztürk Birima, Ipek Kazancoglub, Sachin Kumar Manglac, Aysun Kahramand, Satish Kumareg, and Yigit Kazancogluf' used topic modeling to identify fake reviews in the study article. Topic modeling can catch fake review language trends, according to the authors. The research collected real and fake TripAdvisor hotel reviews. The authors then identified review topics using Latent Dirichlet Allocation (LDA), a prominent topic modeling approach. They employed Support Vector Machines (SVM), Random Forest (RF), and Naive Bayes (NB) to categorize reviews as authentic or fraudulent based on the subjects detected. The suggested technique detected fake reviews accurately, with SVM being the best classifier. In our dataset, we tried the Random Forest.

In the 'Journal of King Saud University – Computer and Information Sciences', 'Arvind Mewada and Rupesh Kumar Dewang' have identified in their paper 'Research on false review detection Methods: A state-of-the-art review' the relation between fake reviews and groups of fake reviewers. They analyzed and pointed out the existing research problems in data acquisition, false feature design, and recognition method design to suggest future research on false review detection. The development of a model for detecting spam reviews is based on various review features and makes predictions on the review text label. The labeled data-based detection method is the most commonly used approach, which involves using supervised machine learning to classify reviews into fake and non-fake categories. Existing methods use crowdsourced and commercially labeled data, and commonly used classifiers include Support Vector Machine, Naive Bayes, Logistics Regression, and Decision Tree. The classification methods used can be based on the dimensional characteristics of reviews or can combine the reviewer's characteristics and their relationship. In our dataset, we tried the Logistics Regression.

Methodology

An illustration of our proposed framework is detailed in **figure 1.**

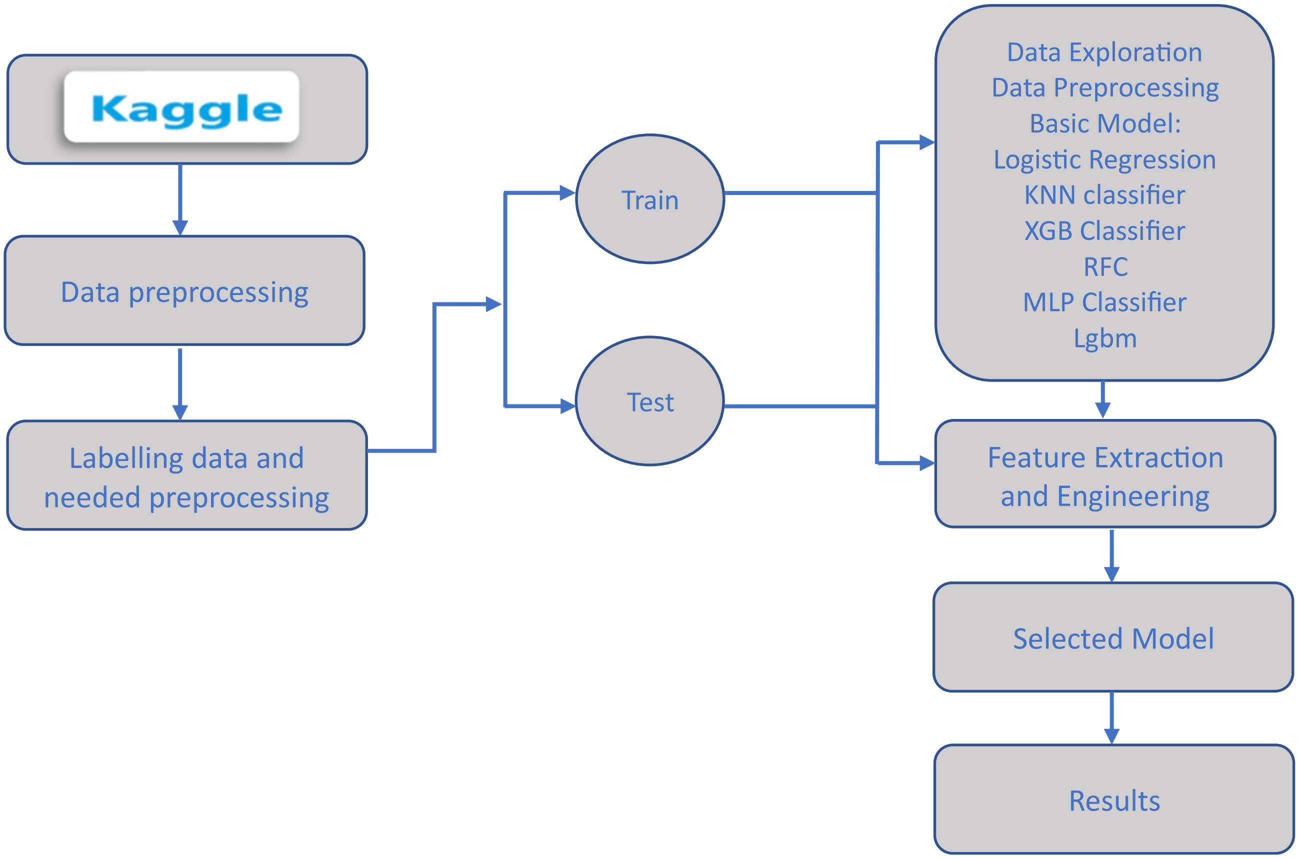


figure 1

Description of the Dataset

This dataset contains 2 files as you see in the below tables
The first file "books_rating" data contain feedback about 3M user on 212404 unique books the data set is part of the Amazon review Dataset it contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014.

Features	**Description**
id	The Id of the book
Title	Book Title
Price	The price of the book
User_id	Id of the user who rates the book

Profile Name	Name of the user who rate the book
review/helpfullness	Helpfulness rating of the review
Review/score	Rating from 0 to 5
review/summary	The summary of a text review
review/text	The full text of a review

The second file “Book_data” data contains details information about 212404 unique books it file is built by using google books API to get details information about books it rated in the first file.

Title	Book Title
describe	Description of the book
authors	Name of the book authors
image	Url for book cover
Preview link	Link to access this book on google books
publisher	Name of the publisher
Publish date	The date of the publish
Info link	Link to get more info about the book
categories	Genre of the books
Rating count	Average rating for the book

First in our data we have done a data preprocessing workflow that involves cleaning, transforming, and labeling data for use in machine learning applications. The initial steps involve dropping null values and merging data frames to create a single, unified dataset. Next, the data is filtered based on a specific range of published dates to include only relevant data (2005-2013)

Data Preprocessing:

Before labeling, the data is prepared for analysis by converting the data types of specific columns, such as authors, review helpfulness, and summary, to strings. The review text and review score are used to label the data using the count vectorizer method. Finally, the labeled data is saved as a CSV file for further analysis.

Data Labeling:

Since our data doesn't have any labeled feature that will help us train a machine learning model in order to identify if the review is fake or true. To fix this, we took advantage of one of our research papers as we used the same dataset that they used in order to label their data. We named the labeled data used as a corpus. First, we selected only three features from our dataset and theirs which were rating, and review/text as for the verified purchase feature we dropped this column since it doesn't exist in our dataset. We faced several challenges to find this feature from other datasets and to relate it to the book ids we had, one of the options was to web scrape Amazon reviews web page, it was not easy to do this task.
After several trials to find a verified purchase in order to link it to our dataset, we went with the option of training the model without having this feature.
Referring to the research paper, using this model on their dataset they had an accuracy of 62% when labeling the data 1 and 0 without the verified purchase feature, as for the model trained and fitted while the dataset includes the "verified purchase" they had an accuracy of 81%.
So we used the model with 62% accuracy since our data doesn't include "verified purchase".
The impressive thing is that while training the log reg on their data we only had an accuracy of 65%, and after implementing the LazyClassifier we found that the light gbm had the highest accuracy 68% while taking into consideration the time taken for execution (5.83 s).

After selecting light gbm as the model to use we had to do some hyperparameter tuning in order to find the best hyperparameters to our model. So we used GridSearchCV(): it helps in finding the best hyperparameters for a machine learning model by exhaustively searching through a specified set of hyperparameters using cross-validation. Hyperparameters are model parameters that cannot be learned during the training process, but are set before training and affect the performance of the model.

GridSearchCV() helped us find the best hyperparameters in our model that were learning_rate = 0.1, max_depth = 7 , n_estimators = 200
Having a learning rate of 0.1 means that at each iteration of the training process, the model's parameters are adjusted by a step size of 0.1 times
the gradient of the loss function with respect to those parameters. In other words, the model will update its weights and biases by taking relatively large steps towards the minimum of the loss function.

Having a maximum depth of 7 for a decision tree means that the tree can have at most 7 levels, including the root node. Each level corresponds to a feature that the tree can split on, and the deeper the tree, the more complex the decision boundaries it can create.
A maximum depth of 7 can be considered a moderately complex tree, which can capture some non-linear relationships in the data but is less likely to overfit compared to a deeper tree.
Having n_estimators set to 200 means that the ensemble model is composed of 200 individual decision trees. Each tree is built using a random subset of the training data and features, in a process known as bagging or bootstrap aggregating.

The ensemble method combines the predictions of all individual trees to make the final prediction.
The data is then labeled using a machine learning model, specifically an LGBM classifier, trained on a research paper. The model is used to label the reviews as fake or real, and then the labeled data is saved as a CSV file.

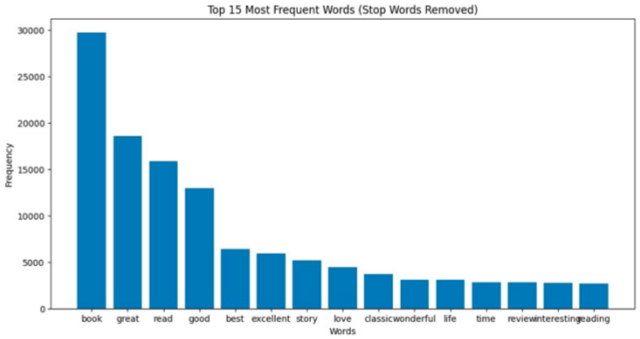
Data Subset:

We dropped the columns we don’t need, “the review text”, and the “preprocess_text” in order to avoid data tempering. We filtered the data “the review time” according to the published date, so since the review could be on an old version of the book which will have a review time before the new publishing of the book, we disregarded all the reviews that are before the published date.
From the “helpfulness” feature we chose the first score that identifies the number of likes for each review, which will indicate the number of users that found the review helpful.
We dropped all the null values and we applied CountVectorizer for the title to see if it has any effect on the nature of the review.
Then, we split our dataset into train and test (80% - 20%) using stratify sampling since we have unbalance in our data.

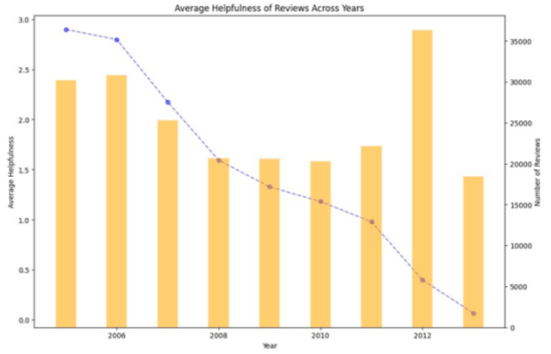
Data Exploration:

We applied data exploration on the train data:

- We could see that the most repeated words have positive feedback for example word "good" and "best" are repeated more than 5000 times in our dataset. In addition, we could see that other "positive" words were dominating the 15 most frequent terms while taking into consideration the removal of stop words that could affect the ordering. The other "positive" words repeated were 'love', 'wonderful', and 'excellent'.



- Throughout the years a decrease in average of helpfulness is noticed, which might indicate that the number of unhelpful reviews increased or people stopped voting on the recent reviews during this period.

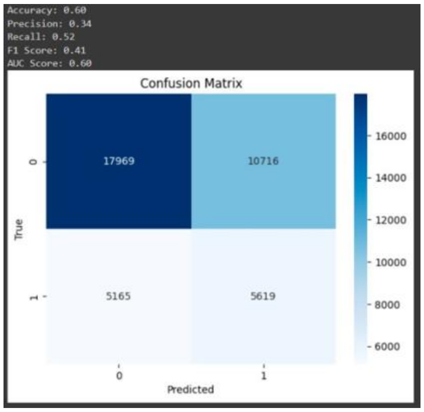


Data Preprocessing:

We applied label encoder to the authors feature and minmaxscaler for the all date-time features and helpfulness feature. The mentioned steps were applied on the test data also.

Basic Models:

We applied several models as basics to detect the accuracy such as Logistic Regression, KNN, Random Forest Classifier, and MLP Classifier that ranged from 50% to 70%. However, we chose the best model which was the Light gbm classifier which gave the evaluation metrics in the screenshot.



```
Accuracy: 0.60
Precision: 0.34
Recall: 0.52
F1 Score: 0.41
AUC Score: 0.60
```

Feature selection and engineering:

As a first step, we dropped the authors column, as mentioned in the literature review, they didn't have an authors column. As a result the accuracy score maintained 60% but the auc score increased 1%.

Then, we increased the max_number of features to 200 as the most repeated words in the review and removing the stop words, but it didn't affect the accuracy. Therefore, we reverted the features to 150.

Moreover, we selected the top 10 most important features which are 'publishedDate_day', 'publishedDate_month', 'publishedDate_year', 'review/time_day', 'review/time_year', 'review/time_month', 'review/helpfulness', 'love', 'review/summary_book', and 'death'. As a result, the accuracy increased to 73% and the AUC score to 61%.

In addition, we created a new feature that measures the difference between the published date and the review time named 'difference'. It was added to the previous feature engineering steps. As a result, the accuracy increased to 76% and the AUC score to 59%.

Results:

To evaluate the effectiveness of the methodologies used in this research concerning the identification and detection of Amazon's fake books' review different experiments were carried out on the data by applying machine learning algorithms and the results obtained were promising, efficient and reliable. Two metrics were used for evaluation. The metrics are accuracy and AUC. The confusion matrix was also used to measure the performance of our selected classifier on our test data.

Conclusions and Recommendations:

We can conduct that we were able to detect 76% out of 68% from the original data. Moreover, feature engineering and feature selection like the difference feature that represented the difference of time between the published book and review date improved the model accuracy.

As recommendations for improving the accuracy of our model, we could have use sentiment analysis in order to predict if a review is positive, negative or neutral. Using this feature, it could have enhanced the performance of the model chosen while predicting Amazon’s fake books’ review. Moreover, webscraping on Amazon website would have been helped in identifying new features that would improve the performance of the chosen model. For example, if we had verified purchase column, published time, and review time in order to extract the duration of the review as mentioned in the literature review above, it could have increased the accuracy in labeling.

References:

Kaushik Daiv1, Mrunal Lachake2, Prathamesh Jagtap3, Srishti Dhariwal4, Prof. Vitthal Gutte5. An Approach to Detect Fake Reviews based on Logistic Regression using Review-Centric Features. In International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 06 | June 2020, Page 2107 - Page 2112, ISO 9001:2008 Certified Journal.

‘SuleO¨ztürk Birima, Ipek Kazancoglub, Sachin Kumar Manglac, Aysun Kahramand, Satish Kumareg, and Yigit Kazancogluf. Detecting fake reviews through topic modelling. In Journal of Business Research 149 (2022) 884–900.

Arvind Mewada, Rupesh Kumar Dewang. Research on false review detection Methods: A state-of-the-art review. In Journal of King Saud University – Computer and Information Sciences 34 (2022) 7530–7546.

Tommaso Fornaciari and Massimo Poesio. Identifying fake Amazon reviews as learning from crowds. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 279–287, Gothenburg, Sweden, April 26-30 2014. c 2014 Association for Computational Linguistics.

https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

https://www.kaggle.com/datasets/kritanjalijain/amazon-reviews