Influence of Fake Reviews on Quality Amazon Product Recommendation

Nisarga Kadam **Nicholas Cheung** **Rei Imai**

Abstract

As the number of products available on Amazon continues to increase, the role of recommendation systems in filtering out irrelevant or poor-quality products by providing users with personalized recommendations is becoming more important. The semantically rich user reviews of a product are the key to providing good recommendations. However, as more products begin to be swarmed with fake reviews, the reliability of recommendation systems starts to falter. In this paper, we attempt to quantify the effect that fake reviews have on recommendation systems. First, we train an ensemble model to classify reviews as real or fake. Next, we run our model on the Amazon Reviews 2023 Electronics dataset to obtain a subset of the dataset containing only real reviews. Finally, we evaluate the original dataset and our newly constructed dataset on the sequential recommendation task using BLaIR, a pre-trained recommendation model and compare results. Our results show that fake reviews have a significant impact on the quality of recommendations.

1 Introduction

As the volume of content online continues to exponentially increase, the need for systems to efficiently and effectively filter out irrelevant content for users becomes increasingly important. On the web, recommendation systems play a key role in addressing this challenge by leveraging information such as user data, browsing history, content information, and metadata to provide personalized content that aligns with a user's preferences. For platforms such as Amazon, product reviews provide another valuable source of information for enhancing recommendations.

While product description and product metadata provide basic details, product reviews offer semantically rich insights that allow for a more nuanced picture of the quality of a product. By incorporating product reviews, recommendation systems are able to provide recommendations that go beyond the often biased perspectives of what the seller or manufacturer of a product has to say about their product. As a byproduct, however, the use of product reviews also further adds to the everlasting challenge of trust in recommendation systems.

User trust is essential for the success of recommendation systems and is at the core of recommendation system development. If a recommendation system suggests irrelevant products, users lose trust in the system and will stop using them. The integration of product reviews only further complicates this question by introducing the questions of "who is writing the review" and "should we be trusting them?"

Fake reviews are reviews that attempt to positively or negatively influence the perception that a user has about a product. While it is nearly impossible to definitively measure the number of fake reviews on any one product, fake reviews undoubtedly have some influence on the reliability of recommendation systems. There have been multiple attempts to create classifiers that are able to detect whether a review is real or not, but these attempts do not tie this problem back to the reliability of recommendation systems.

In this work we make the key contribution of quantifying the impact that fake reviews have on recommendation systems. By exploring this contribution, we hope to advance the field of recommendation system development and offer an alternative way to advance the field through initial data transformation.

2 Previous Research

There has been a variety of research conducted in both fake review detection and recommendation system optimization.

(Fornaciari et al., 2014) worked to identify fake reviews from real reviews on Amazon. The study raised the question: is there an effective system to identify fake reviews? The authors hypothesized that by combining a dataset of fake reviews and using a method to eliminate weak labels from their study, the learning from crowds algorithm, they can create a stronger system that identifies fake reviews.

The study created a corpus that includes book reviews that are real, and fake, and reviews that are neither certainty. They identified features that would help identify a "fake review". The features included a suspect book, a book that is so classic that a review does not make a difference to the purchase of books by Conan Doyle or Stephen King. Companies pay people to write a specific number of reviews over a short period, this resulted in another feature: cluster, identifying reviews written over a short period of 3 days. Nicknames are another feature, as individuals who write fake reviews may not use their real names. Finally, another feature was if the reviewer had purchased the book. Additionally, they used two different approaches: the silver standard used the corpus and the features described above, and the gold standard consisted of a corpus with reviews by Sandra Parker who claimed she had been paid to write them and other reviews of a book where the author admitted he purchased reviews.

The authors conducted two experiments: they used majority voting where if the 3 or 4 deception clues were found in a review, they could be categorized as a fake review, true otherwise. The other experiment was the learning from crowds algorithm where they calculated the probability of each review being true by calculating if each deception clue was dependable. As in prior text classification studies, the authors used a Support Vector Machine as their chosen model. Learning crowds had better accuracy, precision, and recall; whereas the majority voting performed well in precision but the recall was lower than the baseline recall. Above all, combined with the gold standard, the learning crowds approach performed better overall.

(Hou et al., 2024) introduced BLAIR (Bridging Language and Items for Retrieval and Recommendation) as well as the Amazon Reviews 2023 dataset. They used the Amazon Reviews 2023 dataset to train a series of sentence embedding models using a contrastive objective that links user reviews with item metadata. In this way, BLAIR is able to use the context obtained from reviews to filter out items with false descriptions or poor functionality and provide users with a recommendation of items of high quality. They used three different tasks to evaluate their model: sequential recommendation, conventional product search, and complex product search. Sequential recommendation involves predicting the next item of interest from a user's product interaction list. Conventional product search involves labeling the right products as relevant given a specific short list of requirements. Complex product search is similar to conventional product search except the query provided isn't necessarily short and may contain irrelevant information. BLAIR was shown to achieve the best performance for all of these tasks.

In this study, we utilize the methods presented in both of these papers and combine them together to bridge the gap between fake review detection and developing recommendation systems

3 Data

There are two different datasets we are using for building our recommendation system. The first dataset is the fake review dataset which contains 20,000 computer-generated fake reviews and 20,000 human-based real reviews (Salminen et al., 2022). Each review is categorized into 5 distinct genres that appear on Amazon.com: Kindle Store, Books, Pet Supplies, Home and Kitchen, and Electronics. Each item in the dataset consists of 4 fields: category of the item, numeric rating of 1 to 5, text label of CG (computer generated) or OR (original review), and the text body of user review. This dataset is used to train the ensemble model for classifying fake or real reviews.

The second dataset is the Amazon Review 2023. This is a publicly available and large-scale dataset published by McAuley Lab along with the research (Hou et al., 2024). It contains 570 million user reviews on 48 million products that are posted on Amazon.com. Reviews are collected from a wide range of time periods such that the data is extracted from 1969 to 2023. Reviews are grouped by 34 unique categories, and for each category, a json file for user reviews and item metadata are provided. This dataset is being used on the fake review classifier model which we had trained from the prior dataset. From this model, we are going to extract the real reviews from Amazon Reviews 2023, and then pass that dataset to BLaIR in order to compare the result from the original dataset that is also passed to BLaIR.

3.1 Data Pre-processing

Both the fake review dataset and the Amazon Review 2023 dataset are being processed prior to the training of the classifier model. For both datasets, the same preprocessing steps are applied. First, we extracted the text field of 100,000 user reviews under one category from the provided file and turned it into a .txt file. Second, we used the Natural Language Toolkit on Python to tokenize the text that is saved in the .txt file. After tokenizing the review text, we normalized it by removing punctuations, stopwords, and emojis from the tokens.

User reviews that had been extracted into a .txt file are limited to 100,000 from the entire entries due to their large data size. Processing millions of reviews will require a significant amount of time during .txt conversion and tokenization. We have randomly chosen this number which can be sufficient to train our model and still be viable. Although emojis may be an interesting feature to find a connection between fake reviews detection and product recommendation, we chose to focus on text-based reviews for our study.

3.2 Statistics Summary

Table 1: Data Statistics Summary for 100000 samples

Metric	Original	Processed
Number of tokens	8,183,100	3,655,112
Vocabulary size	98,514	81,139
Type to token ratio	0.012	0.022

4 Methods

In order to quantify the impact that fake reviews have on recommendation systems, we first need to train a model to detect fake reviews. Once we have a working model, we run it on our 100,000 Amazon Reviews sample dataset and create a new dataset from that dataset that contains only reviews classified by our model as real. After obtaining this new dataset, we then run our recommendation system on both datasets and compare results.

4.1 Fake Review Detection

We chose three different models to include in our Voting Classifier, the center of our ensemble learning model. We chose to use “hard voting” which counts the predictions made by our model and the class that gets the most votes is the final prediction, whether the prediction is real or fake review. To combine predictions: we chose a Bagging Classifier, a Decision Tree Classifier, and a Logistic Regression.

A bagging classifier breaks up our data into smaller subsections and then combines their results, reducing bias and overfitting. We used a Support Vector Classifier for our bagging classifier. A decision tree is useful because it can mirror and split into trees in the same ways humans make decisions. Logistic regression is not as complex; however, it is efficient at predicting unknown data, which is our goal. The Bagging Classifier covers any overfitting, the decision tree classifier can reflect human decision-making patterns and the logistic regression is efficient at making predictions, the Voting Classifier uses all of the results of these models to create more refined predictions on whether a review is real or fake. Our approach uses a total of four models to identify if a review is real or fake. Our approach is more sophisticated as it accounts for the bias of all models but refines their predictions to increase the chances of correctly predicting a real review.

4.2 Impact on Recommendation System

We used the Bridging Language and Items for Retrieval and Recommendation (BLaIR) model introduced by (Hou et al., 2024) to recommend Amazon products based on our reviews. As mentioned previously, BLaIR is a series of pretrained sentence embedding models that aligns user reviews with item metadata. By doing so, it is able to output items with similar user review sentiment and metadata.

First, we use all reviews, real and fake unfiltered, to recommend Amazon products using BLaIR. Then, we use our newly crafted dataset containing only real reviews to recommend Amazon products using BLaIR. Our goal is to identify if there is an improvement in product recommendations after the fake reviews have been filtered out.

5 Baseline

5.1 Fake Review Detection

To identify the significance of our metrics, our baseline includes a Support Vector Classifier as well as using real and fake reviews to make Amazon electronic product recommendations using BLAIR.

Using baselines allows us to be more transparent in our research so we have clear metrics from the start to improve upon. Baselines allow us to further understand how fake reviews can promote inaccuracy in product recommendations, ultimately leading to unhappy users. Our including fake reviews either significantly boost or hinder a product's sales. As shown in Table 1, our Support Vector Classifier implemented with Bootstrap Aggregating, did not perform well and was only predicting 36% of our records as their actual class. As a result, we recognized that our baseline model was flawed despite Support Vector Machine models being a commonly used learning technique for Natural Language Processing tasks. We used bootstrap aggregation to create smaller subsets of our data and then ran the Support Vector Classifier on these smaller subsets. Our goal was to reduce noise as large datasets are likely to have a lot of variability in their data. Additionally, this problem increases when our points of data are tokenized words, not floating point or integer data. We ultimately realized we were unable to solve the problem of noise given the low F1 score, and our predictions on our Amazon reviews dataset would likely be inaccurate. As Suport Vector Clasifiers are likely to be more specific, they can also be computationally expensive. We recognized that Support Vector Classifiers were not an efficient model to use when scaling as it would reduce the generalizability of a project if it became difficult to scale to large datasets. In our field of research, datasets of reviews can include millions of data, and if our model may not be optimal in cases of resources, a pivot was needed to ensure that our models performed well in our metric of choice as well as being able to generalize well to larger datasets. After this, we decided to implement an ensemble learning model which would include multiple models to reduce noise, overfitting, and scale well to larger datasets while simultaneously predicting more classes correctly.

5.2 Recommendation System

The baseline for our Recommendation System is the original 100,000 Amazon Reviews dataset. Our goal is determine whether or not fake reviews have a noticeable impact on recommendation systems and so we want to compare the results that we get by running our model on the original dataset with all reviews, with the results that we get by running our model on an augmented dataset containing only real reviews. If there does exist a noticeable difference, we have shown that fake reviews have a considerable impact on recommendation systems.

6 Results + Evaluation

6.1 Fake Review Detection Model

To ensure that our model was predicting accurate real reviews, we implemented 4 different classifiers: Bagging using a Support Vector Classifier, a Decision Tree Classifier, Logistic Regression, and ultimately, an Ensemble Learning Classifier. We chose to use the F1 score as our metric because our goal is to reduce false positives and false negatives. We began with Bootstrap Aggregation using the SVC; however, as depicted by Table 1, the F1 score was significantly low. The low score indicates that the model is not efficient at correctly identifying real or fake reviews as their correct label. As visualized, the Ensemble Learning F1 score was in between the scores of the Decision Tree and Logistic Regression. It is significant for us to use Ensemble Learning as it ensures that the logistic regression model is not overfitting. The Ensemble Learning model took a weighted average and 72% of all reviews were correctly identified as either real or fake. Our F1 score shows a balanced value as our goal is to ultimately generalize our results and implement our model to larger datasets of reviews.

6.2 Comparing Fake Review Detection F1 Scores

Table 2: Comparing F1 Scores

Model	**F1 Score**
Support Vector Classifier (Bagging)	0.362
Decision Tree	0.707
Logistic Regression	0.795
Ensemble Learning	0.720

In Table 2, the accuracy for our Ensemble Learning model is 73%, which indicates that 73% of all predictions are correct. Since the values of the F1 score and accuracy are similar, we recognize that our model is predicting accurately a majority of the time and overall balanced in its class predictions; this means that our model is not significantly predicting one class over the other. The F1 score directly shows how well our model is predicting both classes of reviews, real reviews, and fake reviews. An F1 score of 72% is satisfactory; however, the model can still be refined through extensive preprocessing. Feature engineering metadata would provide more parameters for the model to recognize the patterns. Furthermore, adding more classifiers to ensemble learning would further refine how balanced the true positives and false negatives are. Although these results favor generalization, to better apply our model, we could use feature engineering, and larger datasets of fake reviews to better understand writing patterns and have better-performing metrics.

6.3 Ensemble Learning Model Classification Report

Table 3: Ensemble Learning Classification Report

Index	Precision	Recall	F1-Score	Support
0	0.68	0.94	0.79	419
1	0.88	0.51	0.65	379
Accuracy			0.73	798
Macro avg	0.78	0.72	0.72	798
Weighted avg	0.77	0.73	0.72	798

6.4 Newly Acquired Dataset

Using our Fake Review Detection Model, we classified every review in our 100,000 size sample of the Amazon Reviews Dataset as real or fake. Our model detected 80,701 of the reviews as fake and the remaining 19,299 reviews as real. Using only the 19,299 reviews that were labeled as real reviews, we created a new dataset to run BLaIR on.

6.5 BLaIR Performance

Table 4: Performance comparison on different Electronics datasets.

Datasets	R@10	N@10	R@50	N@50
Real + Fake	0.0189	0.053	0.0095	0.0168
Real	**0.0667**	**0.1889**	**0.0316**	**0.0563**

The task used to evaluate BLaIR's performance on the two datasets was the sequential recommendation task. Given the interaction sequence of one user $\{i_1, i_2, \ldots, i_l\}$ the task is to predict the next item of interest i_{l+1}. i represents the sentence corresponding to the metadata of a particular item, and fortunately, included in the metadata for the Amazon Reviews 2023 dataset is a list of interactions with other items, so we can determine relevant i_{l+1} items.

The metric of evaluation used for this task is the Recall@10, Recall@50, normalized discounted cumulative gain (NDCG) @10, and NDCG@50. Recall@N is the fraction of relevant items that the recommendation model recommended that were in the top N items of interest. NDCG@N is similar to recall, but compares the ranking that the model produced to the ideal order of relevant items.

BLaIR's performance on the two datasets on this task can be seen in Table 4. On all four metrics, BLaIR performed considerably better when ran on the dataset containing only real reviews compared to when ran on the entire dataset.

7 Conclusion

The aim of this project was to create more accurate Amazon product recommendations to increase user satisfaction rates and trust within online marketplaces. As online shopping has skyrocketed over the last decade, whispers of transparency about the products have become a point of contention for users as well as an avenue for influencers to increase their market cap. Social media is littered with product reviews that have historically hidden partnerships with brands and led to mistrust between consumers and retail giants. As companies continue to incentivize influencers and celebrities to positively review products, as well as pay individuals to write fake positive reviews, the value of a real review from a layman has never held more value. Real reviews from real people are vital and they should be a majority of the reviews we encounter. To ensure this, we decided to use a dataset of fake reviews to identify which reviews from Amazon were likely to be real or fake reviews. When creating product recommendations, it is imperative that companies perform testing and modeling with the same goal as real reviews of product that are honest about product quality will only lead to better sales and good faith with the consumer.

We initially tried to use a Support Vector Classifier and Bootstrap Aggregating to refine our class predictions; however, this model returned a poor F1 score, our metric of choice. We turned to other models that could reduce overfitting, scale better to large datasets, and are efficient at class predictions. We combined the results of our Decision Tree Classifier, Logistic Regression, and Support Vector Classifier through an Ensemble Learning method called the Voting Classifier. Our F1 score of .73 is significantly better than our F1 score of the Support Vector Classifier, which served as our baseline model, as .36. With our BLaIR recommendation system, we used real and fake reviews as our baseline and then subsequently only used real reviews to identify if our model was working well. As shown in our results above, we were able to obtain better results when we filtered out the fake reviews and used real reviews to recommend Amazon products. Our limitations include that we did not significantly pre-process our data with more metadata. Feature engineering would have helped us to balance precision and recall better since we would have had more data to create refined patterns. Having a sophisticated understanding of BLaIR would have aided us in creating a more comprehensive analysis and

Overall, our work shows that we created a foundation we could build upon to get better product recommendations. As predicted, real reviews from real users were more accurate when used to make better product recommendations. Our project provides an insight into an expansive field of research and we hope to provide companies with an avenue to build their research upon to continue gaining consumer trust in an industry that is significantly influenced by trends.

8 Future Work

Identifying fake reviews is critical for creating better user experiences. Given the rise in prices, consumers are cautious about their spending, and detecting fake reviews would ensure that users can purchase quality products, resulting in consumer satisfaction and restoring consumer trust in marketplaces.

Using complex models and many preprocessing techniques can ensure more correct predictions. A potential avenue would be to understand more writing patterns in regions and combine that with review metadata such as IP address to identify if the review matches the writing patterns of a consumer in that area. Better detection of fake reviews could also help regulate and penalize companies that use fake reviews to increase sales. Some challenges in this field are that speech is constantly changing, a large amount of data collection, and the technical limitations of continuing this work in different languages. As speech is changing constantly, it may result in less accuracy in predicting older reviews. At least on Amazon, a large amount of user data is collected, this raises data privacy concerns as including more user information as features can be violating user privacy. Additionally, there may not be as many libraries and text pre-processing opportunities available in other languages to detect fake reviews efficiently. Navigating these issues will not only enhance user experience but will also ensure user trust.

Additionally, in this study, we evaluate the impact of fake reviews on a single recommendation system model, BLaIR. In the future, we would like to evaluate the impacts that fake reviews have on a variety of other models and see whether or not fake reviews affect different models similarly. Filtering out fake reviews is a critical part of product recommendation and as our economy continues to increase its reliance on online marketplaces, it is vital to ensure we create robust product recommendation systems; this includes feature engineering and methods of filtering out irrelevant reviews to further gain consumer trust and subsequently, increase profitability.

References

- Hou, Y., Li, J., He, Z., Yan, A., Chen, X., McAuley, J.J.: Bridging language and items for retrieval and recommendation. arXiv:2403.03952 (2024)

  Passon, M., Lippi, M., Serra, G., Tasso, C. (2018, November 1). Predicting the usefulness of Amazon reviews using off-the-shelf argumentation mining. ACL Anthology. https://aclanthology.org/W18-5205/

  Tommaso Fornaciari and Massimo Poesio. 2014. 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. Association for Computational Linguistics.

9 Contributions

Nisarga: Fake review detection (baseline, all learning models, results evaluation), previous research

Nicholas: Abstract, Introduction, Previous research, BLaIR

Rei: Previous Research, Data Pre-processing