IDENTIFYING HIJACKED REVIEWS

A Thesis

by

MONIKA MANOHAR DARYANI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	James Caverlee
Committee Members,	Theodora Chaspari
	Patrick Burkart
Head of Department,	Scott Schaefer

May 2021

Major Subject: Computer Science

ABSTRACT

Customers on online marketplaces have to proceed with extreme caution before buying any product as they cannot evaluate it physically. Reviews are crucial metrics to gauge the quality and authenticity of the item. This dependence on reviews has led to a rise in the number of unethical sellers who exploit the review system in e-commerce websites via fraudulent techniques, like fake reviews.

Fake reviews have been an actively researched domain for the last decade as an independent review problem and a behavior pattern recognition problem. While almost everyone is looking around to detect fake reviews, we are looking at a different facet of e-commerce fraud which is called “Review Hijacking” (or “Review reuse” or “Bait-and-Switch review”).

Review hijacking is a new review manipulation tactic in which black-hat sellers “hijack” existing review listings of a product and use them to sell their products with no reviews. These items may be discontinued and unrelated but contain many positive reviews. More favorable ratings lead to better search ranking, make a new product appear well-reviewed and legitimate, and, ultimately, boost sales.

There has been little academic research for this review scam. Hence, we introduce what review hijacking is, the methods used to employ it, challenges to identify, and the impact it has caused. We further find techniques to uncover such cases.

We analyze the extent of this problem by applying various Information Retrieval methods like Boolean Retrieval (BIR), TF-IDF and Topic modeling on Amazon public datasets. Then, we synthetically label our data using Weak Supervision and by swapping the product-review pairs to run supervised learning models. We employ Deep Learning methods like Siamese LSTM and BERT Sentence Pair Classification to detect this e-commerce fraud efficiently on a larger scale.

ACKNOWLEDGMENTS

I thank my thesis advisor, Dr. James Caverlee, for his continuous guidance and support throughout this thesis. His advice and directions have assisted me smoothly carry out all my tasks at every step. This thesis work uses concepts and implementations taught in his classes on Data Mining and Information Storage & Retrieval. His feedback improved my thesis document leaps and bounds.

I would thank Yun He, Parisa, Ziwei Zhu, Eddie W, Majid, and all other members of Infolab for their contributions to this thesis. Parisa's, Yun's, and Eddie's directions on implementation and Majid's tips with the dataset helped me get rolling. Ziwei's review of this thesis document and complete Infolab's output and suggestions for my defense presentation aided the completion of my thesis.

I thank my parents, Manohar Daryani and Lata Daryani, as they have always supported me in all my life's endeavors. Finally, I am grateful to my husband, Harsha Aduri, for motivating me, showing confidence in me, helping bouncing ideas off, and all the cheering up for when I would be feeling down.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Dr. James Caverlee and Dr. Theodara Chaspari of the Department of Computer Science and Engineering at Texas A&M University and Dr. Patrick Burkart of the Department of Communications at Texas A&M University. The dataset used in the thesis work was provided by Dr. Julian McAuley of the Department of Computer Science, University of California at San Diego as a publicly available contribution.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

Graduate study was supported by funding from scholarship and part-time student worker positions from Computer Science Department at Texas A&M University.

NOMENCLATURE

USA	United States of America
ASIN	Amazon Standard Identification Number
BERT	Bidirectional Encoder Representations from Transformers
NLP	Natural Language Processing
CNN	Convolutional Neural Networks
ML	Machine Learning
Masked LM / MLM	Masked Language Model
NSP	Next Sentence Prediction
BIR	Boolean model of Information Retrieval
TF-IDF	Term Frequency - Inverse Document Frequency
BM25	Best Matching version-25
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
IR	Information Retrieval
FTC	Federal Trade Commission
SME	Subject Matter Experts
QA	Question-Answering
NLI	Natural Language Inference
LF	Labeling Function
LDA	Latent Dirichlet Allocation
ROC	Receiver Operator Characteristic
TP	True Positive

TN	True Negative
FP	False Positive
FN	False Negative
TPR	True Positive Rate
FPR	False Positive Rate
AUC	Area Under Curve
AWS	Amazon Web Services
TSV	Tab-Separated Values
JSON	JavaScript Object Notation
ID	Identity documentation
NER	Named Entity Recognition
MTurk	Amazon Mechanical Turk
GloVe	Global Vectors
SOTA	State Of The Art
GPU	Graphics Processing Unit
TPU	Tensor Processing Unit

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

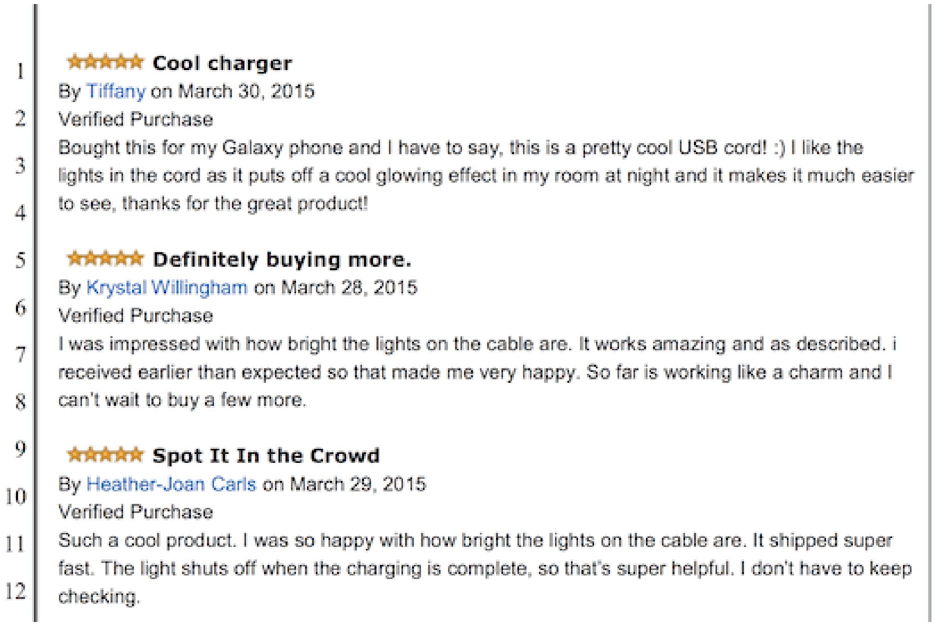
LIST OF TABLES

TABLE Page

1. INTRODUCTION

With the spread of the internet, online e-commerce marketplaces have become ubiquitous and indispensable for buying and selling products globally. They have contributed more than 50% of all e-commerce sales and accounted for $1.7 trillion of the world economy in 2019 [5]. Currently, there are more than 150 online marketplaces, ranging from giants like Amazon, eBay, Rakuten, Alibaba, and Walmart to niche ones like Bonanza, Fruugo, and eBid. While many of these marketplaces function globally, some are region-specific, such as Poshmark in the USA, PayPayMall in Japan, Flipkart in India, and Taobao in China. Some marketplaces are product category-specific, such as BestBuy for electronics, Etsy for Fashion & Gifts, and WayFair for Homeware & Furniture. Indeed, one can buy about nearly everything online, ranging from groceries from Instacart to weapons from GunBroker.

1.1 Consumer Reviews



★★★★★ **Cool charger**
By Tiffany on March 30, 2015
Verified Purchase
Bought this for my Galaxy phone and I have to say, this is a pretty cool USB cord! :) I like the lights in the cord as it puts off a cool glowing effect in my room at night and it makes it much easier to see, thanks for the great product!

★★★★★ **Definitely buying more.**
By Krystal Willingham on March 28, 2015
Verified Purchase
I was impressed with how bright the lights on the cable are. It works amazing and as described. i received earlier than expected so that made me very happy. So far is working like a charm and I can't wait to buy a few more.

★★★★★ **Spot It In the Crowd**
By Heather-Joan Carls on March 29, 2015
Verified Purchase
Such a cool product. I was so happy with how bright the lights on the cable are. It shipped super fast. The light shuts off when the charging is complete, so that's super helpful. I don't have to keep checking.

Figure 1.1: Product Reviews

Unlike physical stores, consumers on online marketplaces cannot evaluate a product in person. Hence, they may find it difficult to trust its authenticity and quality. Many online marketplaces have incorporated a review system to resolve this. They provide an option for the customers to rate and write a review for a product they have used, also add a picture if desired. Figure 1.1 shows some example reviews. These reviews are attached to each item and assist new buyers interested in the product to assess its quality, legitimacy, and reliability. Some platforms also allow users to grade the reviews provided by other users for their usefulness or accuracy.

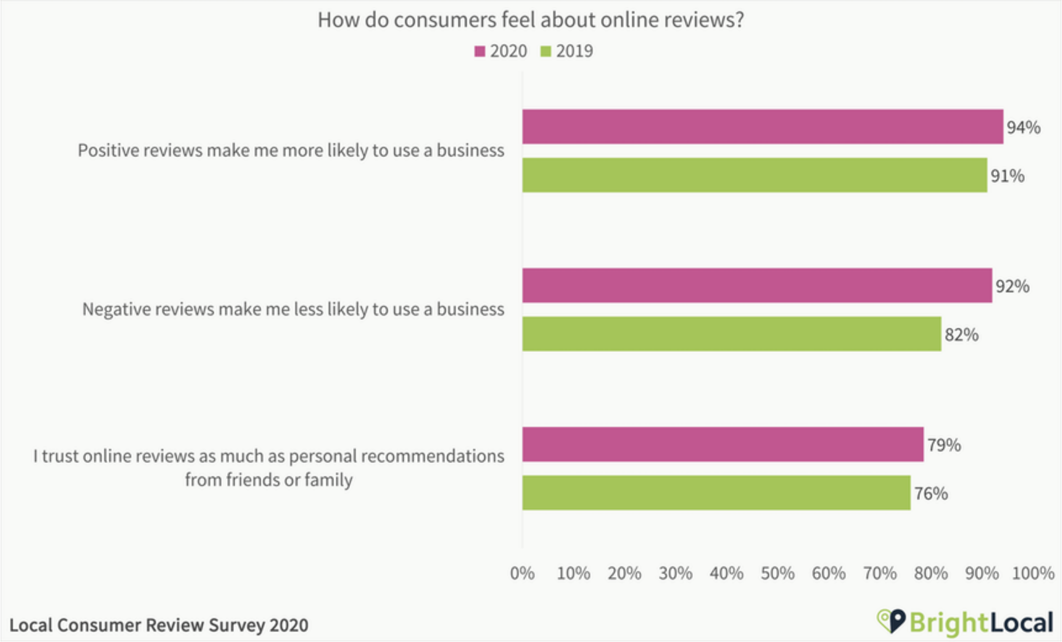


Figure 1.2: Consumer feel about Reviews [6]

These reviews are essential, with surveys providing evidence of their importance in figure 1.2. In 2020, 94% of people admitted that they read online reviews when visiting local businesses. 92% say negative online reviews have convinced them to avoid a product or company [6]. 93% of customers read online reviews before buying a product. 79% people in general and 91% of people in the age group 18-34 years trust online reviews as much as they trust personal recommendations [7]. Thus, these reviews play a crucial role in gaining or losing customer trust and, this importance is only growing every year.

1.2 Review Abuse and Misuse

While reviews serve as a boon to consumers by providing insights into a product even in the absence of personal experience, they are also targets for abuse and misuse. Black-hat sellers exploit the loopholes in the review system and inflate their product sales or hurt their competitors. There have been many reports about review abuse on all major platforms [9] [10] [11] [12]. In 2015, Amazon sued over 1000 people because they posted fake reviews [13]. FakeSpot claims as many as 42% of all reviews on Amazon are unnatural [14]. Even the Federal Trade Commission (FTC) has challenged fake paid reviews [15].

Review abuse takes many forms. Unethical sellers try to increase positive reviews for their products by personally writing positive reviews or enlisting associates. Otherwise, offer a discount, free products, or other compensation to solicit biased reviews. For example, some sellers use Facebook groups to request these incentivized reviews [16] [17]. Some companies employ people to write reviews for service sites [18]. Sometimes unscrupulous sellers also employ fake negative reviews or ratings targeted at competitors to weaken the competition [19]. Sellers hijack listings of other products to collect favorable reviews for their products [3] [11]. They also find methods to delete negative reviews on their products [20]. Indeed, there are full-fledged businesses involved in illegally targeting review systems; one such example is Appsally.

Another form of abuse is click fraud, and it occurs in multiple ways. Bots operate to pretend to be legitimate buyers rating products on e-commerce platforms. Click-farms (Figure 1.3) are organizations that leverage large groups of cheap workers to manually click on paid ads of the product or the product itself to skew the relevance algorithms. These methods either add positive ratings to your product listings or add negative ones to others [21].

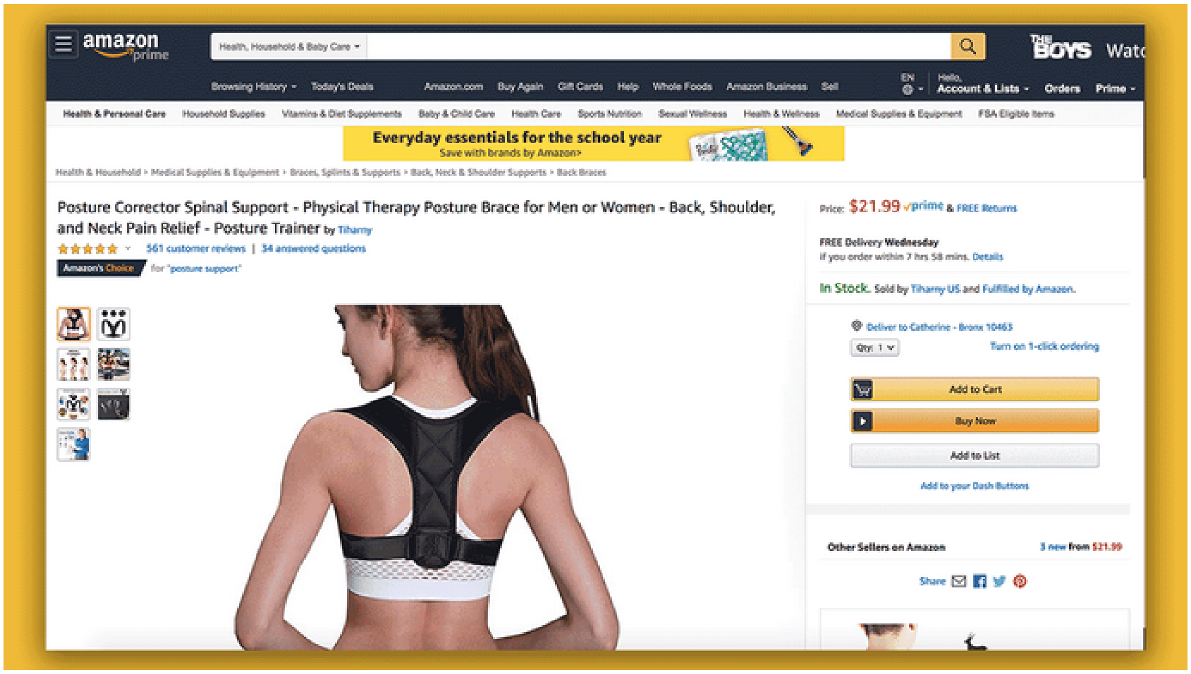
Websites like FakeSpot, ReviewMeta and The Review Index analyze Amazon product reviews and help detect unnatural ones for the buyers.

Figure 1.3: A Click-Farm

1.3 Review Hijacking

As we have discussed, sellers have employed several methods to misuse review systems. We investigate an understudied facet of e-commerce fraud called *Review Hijacking* in this thesis. If you go to a product page and read reviews attached to them, you might find some for an entirely different product. Such off-topic feedback can be the result of this review hijacking. In essence, unethical sellers hijack a legitimate product with many positive reviews to insert their product (which has accumulated no positive reviews). They then reap the ratings "halo" from consumers who assume the new product is highly rated. This review hijacking (or "review reuse" or "bait-and-switch reviews") provides the sellers with an easy way to obtain lots of positive reviews.

Figure 1.4 is an example of review hijacking on Amazon. As seen from the item description, image, and product title in the figure, the product is a posture correction brace. The product seems to have 500+ reviews, a majority of them as 5-star ratings and a "Verified Purchase" tag. As we scroll further and read reviews, we see some are for different products altogether. There are reviews for dish soap, paddles, allergy medications, and so on. Clearly, this figure shows an example of review hijacking.



Posture Corrector Spinal Support - Physical Therapy Posture Brace for Men or Women - Back, Shoulder, and Neck Pain Relief - Posture Trainer by Tiharny

★★★★★ 561 customer reviews | 34 answered questions

Amazon's Choice for "posture support"

Price: $21.99 ✓prime & FREE Returns

FREE Delivery Wednesday
If you order within 7 hrs 58 mins. Details

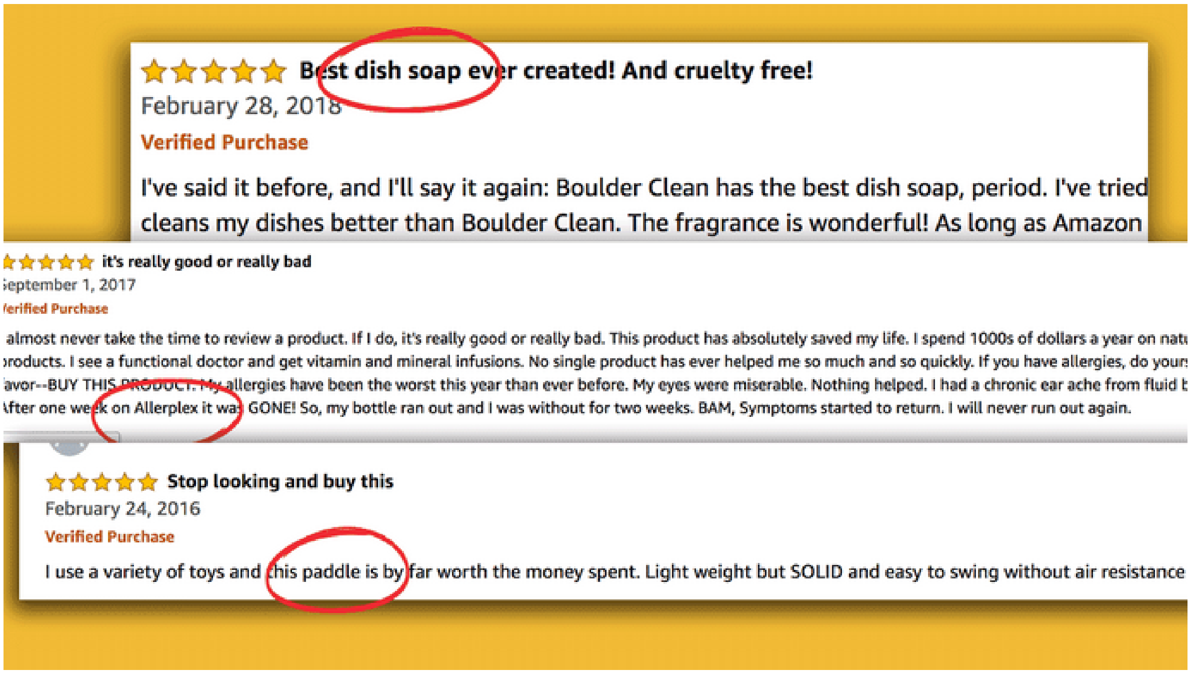
In Stock. Sold by Tiharny US and Fulfilled by Amazon.

Deliver to Catherine - Bronx 10463

Qty: 1 Turn on 1-click ordering

Other Sellers on Amazon 3 new from $21.99

A posture correction brace found on Amazon in early August.



★★★★★ Best dish soap ever created! And cruelty free!
February 28, 2018
Verified Purchase
I've said it before, and I'll say it again: Boulder Clean has the best dish soap, period. I've tried
cleans my dishes better than Boulder Clean. The fragrance is wonderful! As long as Amazon

★★★★★ it's really good or really bad
September 1, 2017
Verified Purchase
almost never take the time to review a product. If I do, it's really good or really bad. This product has absolutely saved my life. I spend 1000s of dollars a year on natu
products. I see a functional doctor and get vitamin and mineral infusions. No single product has ever helped me so much and so quickly. If you have allergies, do your
avor--BUY THIS PRODUCT. My allergies have been the worst this year than ever before. My eyes were miserable. Nothing helped. I had a chronic ear ache from fluid b
After one week on Allerplex it was GONE! So, my bottle ran out and I was without for two weeks. BAM, Symptoms started to return. I will never run out again.

★★★★★ Stop looking and buy this
February 24, 2016
Verified Purchase
I use a variety of toys and this paddle is by far worth the money spent. Light weight but SOLID and easy to swing without air resistance

A selection of hijacked reviews found on the listing for the posture correction brace.

Figure 1.4: Hijacked Reviews on the Listing for the Posture Correction Brace [3]

1.4 Research Goal and Challenges

While this type of review abuse has been recognized anecdotally in the press [3], [11], there has been no structured research to date. Therefore, we aim to conduct the first systematic investigation for review hijacking and propose methods to identify it. Concretely, we explore unsupervised and supervised machine learning methods to detect them. We focus on one online marketplace – Amazon – for which there is considerable anecdotal evidence of review hijacking. We then evaluate the proposed methods to analyze their effectiveness and identify open research avenues for ongoing review hijacking detection.

Thus, our research goal consists of the following four main components:

1. **Study:** Conduct the first systematic investigative study of review hijacking. (Chapter 2)
2. **Explore:** Explore and experiment with unsupervised and supervised machine learning techniques. (Chapter 4 and 5)
3. **Detect:** Propose efficient methods to detect Review Hijacking. (Chapter 6)
4. **Evaluate:** Evaluate the proposed methods to analyze their effectiveness. (Chapter 6)

Chapter 3 discusses datasets and data analysis for this thesis research. Chapter 7 discusses some future work ideas.

But, there are several challenges to this:

- **Lack of Ground Truth:** First, there are no standard datasets of reviews known to be examples of review hijacking. Hence, we have no labeled data. Supervised machine learning methods require gold labels to train effective models and evaluate prediction accuracy. Even unsupervised models require ground truth for evaluation.

  Toward overcoming this deficiency of ground truth, we take two approaches in this thesis. First, we manually label a subset of reviews by reading the product reviews ourselves. Millions of product-review pairs make this a very time-consuming and expensive process. Second, we propose to generate synthetic data. This type of data generation is faster and cheaper, but we trade-off accuracy for these advantages.

- **Absence of Related Work:** Scholars have been actively researching the concept of fake reviews for the last decade both as an independent review problem [31] [32] and a behavior pattern recognition problem [33]. However, the topic of hijacked reviews, even though substantially reported, has little to no public research both from problem understanding and solution proposal perspectives.

  As a solution, we looked into papers which work for title and text matching system in fake news or essay rating system. We also utilized some resources for text similarity. This is further discussed in section 2.3.

- **Huge Datasets:** We used two publicly available datasets for Amazon reviews, as discussed in chapter 3. Each of them has well over 100 million product reviews providing more than sufficient data to train and test. But this enormous size also makes it very resource-intensive and time-consuming to store, run training and even evaluate performance.

  We used a laptop or desktop with high computing and memory resources for some of these tasks. Majorly, we used cloud computing resources like Google Colab Pro and AWS GPU instances. Google Colab Pro is a hosted Jupyter notebook service running on Google Cloud with built-in support for storage via Google Drive. It comes with built-in support for various machine learning frameworks and provides GPU/TPU support. For more resource-intensive on-demand tasks, we resorted to AWS GPU instances. AWS is a pay-as-you-go service provided by Amazon where we subscribe and rent the required virtual hardware, software, and networking features.

Other methods included dividing the dataset into parts. We split the Amazon review dataset into 1000 parts for trouble-free storage and execution. To efficiently use RAM while reading the data, we used buffering io stream for input. We also trimmed down unnecessary fields in the dataset which we did not use for classification, like verified purchase, vine program, date, and time for review. This removal helped reduce the size of product metadata from 87 GB to 22 GB.

- **Sample Sparsity:** Another challenge is that the number of hijacked reviews is extremely low on e-commerce platforms. In our preliminary study, manual labeling suggests that $<$ 0.01% of all reviews are examples of review hijacking. This tiny quantity creates issues when we try to evaluate our models. Generally, we assess machine learning models using classification accuracy, which does not work effectively with such skewed data. We use AUC under the ROC curve and Precision-Recall curves for the results. We discuss these metrics in section 2.4. We also generate synthetic data in Chapter 5 to overcome this for supervised learning methods.

- **Quality of Reviews and Product Titles:** Not all reviews help detect hijacked reviews, with some of them being too generic or too vague. For example: "Great product! Five stars!" can be for any product. We cannot deduce if the review is for the given product or not. Another example, "This looked so pretty", provides information about some product features, but not the product itself. Some reviews explain the shipping or other generic information. Hence, sometimes it is hard to deduce review hijacking.

  A related challenge is that reviewers sometimes write gibberish or enter random numbers in a review, use other languages (making labeling more difficult), or make typing errors. Again, some feedbacks are short (5-10 words), while others are long (with more than 5,000 words). These also create issues for hijacking detection.

Similarly, there are issues with the product titles the sellers put in. At times the product titles are long and descriptive, sometimes they are short, even one word, and do not tell what the product is. These titles might not be an issue on the actual Amazon site as it has images attached. But when detecting using only text, these titles cause difficulty with review hijacking detection.

- **Seller Techniques:** Finally, the hijackers use very sophisticated techniques too. The review listing may have some reviews (incentivized, fake, or real) related to the given product. Thus, a hijacked review listing can be a mixed bag of reviews for the actual product and the hijacked product.

  A seller may also hijack a listing of a somewhat related product. For example, if the product is “iPhone XS cover” and the review is about “iPhone 5C phone”, it can be difficult for the ML model to find the hijacked review.

1.5 Objectives and Findings

The objective here is to understand how Review Hijacking takes place and its impact. Then, we explore various algorithms to detect such fraud, propose an efficient one and then evaluate its performance.

We have unlabeled data. So, our first approach will be using unsupervised learning methods. We use basic algorithms for information retrieval and web searches like boolean retrieval (BIR), TF-IDF, and BM25. We aim to use the product title as a search query and score the reviews. If the reviews get poor scores, those reviews are labeled hijacked. We also used a text summarizing model such as Topic modeling in this part. When the topic or summary of the review text does not match the product title, we tag it as hijacked. We evaluate the results via eyeballing for correctly classified or not.

Unsupervised learning methods helped to obtain some patterns for the review hijacking. We could see that there are examples of review hijacking in our data. The output of all retrieval methods could retrieve hijacked listings with < 25% accuracy. The results for these unsupervised algorithms were suboptimal as there is little to no information about the word or sentence semantics in the methods we used.

Our second approach will be using supervised learning methods. We attempt to generate synthetic labels for the data using weak supervision, which does not work very well. We then create labels by swapping the product title and the review text of two unrelated products. These unrelated product pairs are obtained by (1) using products of two different unrelated categories and (2) measuring Jaccard distance between product titles of the same class. We used supervised methods Siamese LSTM and BERT fine-tuning to classify this data. For both of them, we input product title and description concatenated as one input, review headline and text as other input, and generate results. We evaluate these models using ROC curve results.

Siamese LSTM provided good precision and ROC-AUC result. But, BERT fine-tuning provided better results with 90% accuracy and AUC. We see that BERT provided the best prediction from all the methods we attempted. We keep in mind that this result is generated on noisy synthetic data and might not be 100% accurate.

2. PRELIMINARIES

In this chapter, we review some background on review hijacking. Then, introduce some notations for the rest of the thesis.

2.1 Methods of Review Hijacking

We begin by examining the methods of review hijacking on Amazon.com. We shall discuss three majorly known techniques: (i) Ship of Theseus; (ii) Variations; and (iii) Global Products.

2.1.1 Ship of Theseus

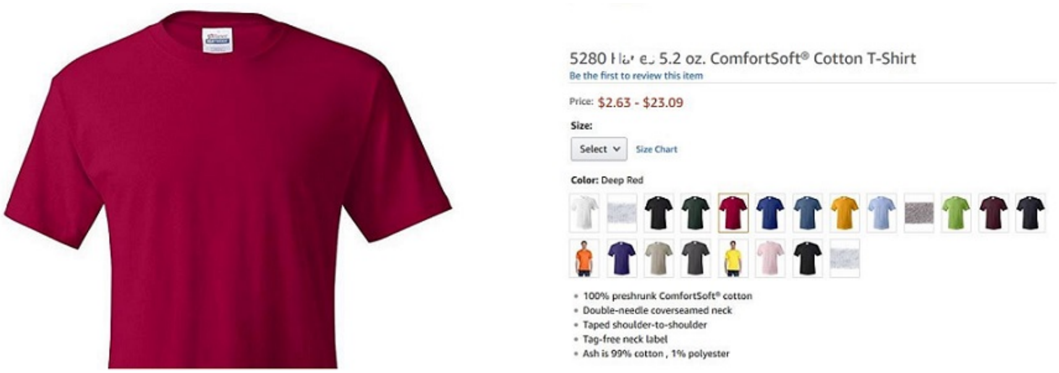
Amazon associates the reviews with a page using a unique Amazon Standard Identification Number (the ASIN) and not the product itself. So, when a seller decides to stop selling a product, the page becomes dormant. The product lists as “Currently Unavailable”, but all the reviews stay attached. An unethical seller can use this opportunity and then update the product’s title, photo, description, and so on. Usually, he makes these changes incrementally to avoid getting noticed. This product can be related to the original product or an entirely different product. The reviews attached to the product appear alongside the new product now, some of which might also be “Verified purchases”, which can carry extra weight for consumers.

This method is similar to the concept of “Ship of Theseus”, a thought experiment about the nature of identity, wherein one replaces the components of an object one by one. It is supposed that the famous ship sailed by the hero Theseus in a great battle was kept in a harbor as a museum piece. As the years went by, some of the wooden parts began to rot and be replaced by new ones. Then, after a century or so, they had replaced every component. The question then is whether the "restored" ship is still the same object as the original. For us, the object is a product listing, and every component is replaced aside from the reviews. For this type of review hijacking, sellers may use any obsolete listing or use one’s listing for this purpose.

For example, Mr. X is a new seller on Amazon and has started selling CD players. After a year has passed, this CD player receives 1000 positive customer reviews, including plenty from verified customers. Then, Mr. X changes the image on this CD player page to a USB cable. He then slowly changes the description, bit by bit, to update it for the USB cable and then changes the title too after few days. These changes result in a product listing for a USB cable that contains reviews of a CD player. He then generates 50 fake reviews for the USB cable to add to the top of the CD player reviews to push them down to the second or third page. This page now appears to the customer as 1050 positive reviews for the USB cable. When customers start buying the USB cable, they might find low quality and add negative reviews. But, positive reviews from the CD player will supersede even if there are 100 negative reviews for the USB cable.

2.1.2 Variations

Amazon lets sellers list variations of a product like dresses or shoes with pattern and color changes (Figure 2.1). These products have their reviews aggregated together. This aggregation makes sense as they are essentially the same product, and you would not want to have to tally up the reviews from each variation to see the entire review picture.



5280 Hanes 5.2 oz. ComfortSoft® Cotton T-Shirt

Be the first to review this item

Price: $2.63 - $23.09

Size:

Select ⌄ Size Chart

Color: Deep Red

- 100% preshrunk ComfortSoft® cotton
- Double-needle coverseamed neck
- Taped shoulder-to-shoulder
- Tag-free neck label
- Ash is 99% cotton , 1% polyester

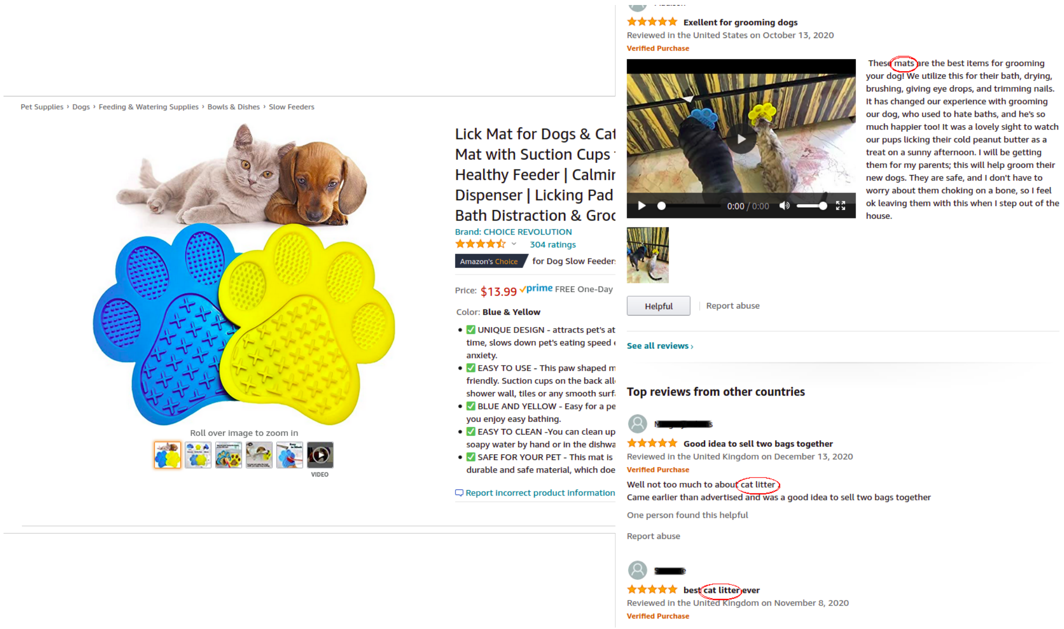
Figure 2.1: Variation Relationship [4]

Some sellers use this feature to unethically list their product as a variation of an unrelated product to aggregate its reviews and high ratings. Sometimes, these unrelated products can also belong to some other seller. Sellers exploit a loophole on the Amazon website as rules for variation are not clearly defined.

For example, a seller may have many positive reviews for their WiFi router. Then, they list an Ethernet cable as a variation of the product and thus group reviews of two essentially different products. They fool customers into believing that even the Ethernet cable is well-reviewed. If customers put in negative reviews, we will have a similar story like the Ship of Theseus approach, where old hijacked positive reviews supersede the newer negative reviews.

2.1.3 Global Products

Sellers on Amazon can sell across multiple countries. Amazon allows aggregating reviews from all countries for a product. Unethical sellers misuse this feature and merge reviews from unrelated products sold in a foreign country to the USA listings. At times these reviews are in some other language, making them tougher to detect.


Pet Supplies › Dogs › Feeding & Watering Supplies › Bowls & Dishes › Slow Feeders
Roll over image to zoom in
VIDEO
304 ratings
Amazon's Choice
Report incorrect product information
Helpful
Report abuse
See all reviews ›
0:00 / 0:00

Lick Mat for Dogs & Cat
Mat with Suction Cups
Healthy Feeder | Calmir
Dispenser | Licking Pad
Bath Distraction & Groc

Brand: CHOICE REVOLUTION

for Dog Slow Feeder

Price: $13.99 ✓prime FREE One-Day

Color: **Blue & Yellow**

- ✅ UNIQUE DESIGN - attracts pet's at time, slows down pet's eating speed anxiety.
- ✅ EASY TO USE - This paw shaped m friendly. Suction cups on the back all shower wall, tiles or any smooth surf
- ✅ BLUE AND YELLOW - Easy for a pe you enjoy easy bathing.
- ✅ EASY TO CLEAN -You can clean up soapy water by hand or in the dishwa
- ✅ SAFE FOR YOUR PET - This mat is durable and safe material, which doe

Exellent for grooming dogs
Reviewed in the United States on October 13, 2020
Verified Purchase

These mats are the best items for grooming your dog! We utilize this for their bath, drying, brushing, giving eye drops, and trimming nails. It has changed our experience with grooming our dog, who used to hate baths, and he's so much happier too! It was a lovely sight to watch our pups licking their cold peanut butter as a treat on a sunny afternoon. I will be getting them for my parents; this will help groom their new dogs. They are safe, and I don't have to worry about them choking on a bone, so I feel ok leaving them with this when I step out of the house.

Top reviews from other countries

[redacted]
Good idea to sell two bags together
Reviewed in the United Kingdom on December 13, 2020
Verified Purchase
Well not too much to about cat litter
Came earlier than advertised and was a good idea to sell two bags together
One person found this helpful
Report abuse

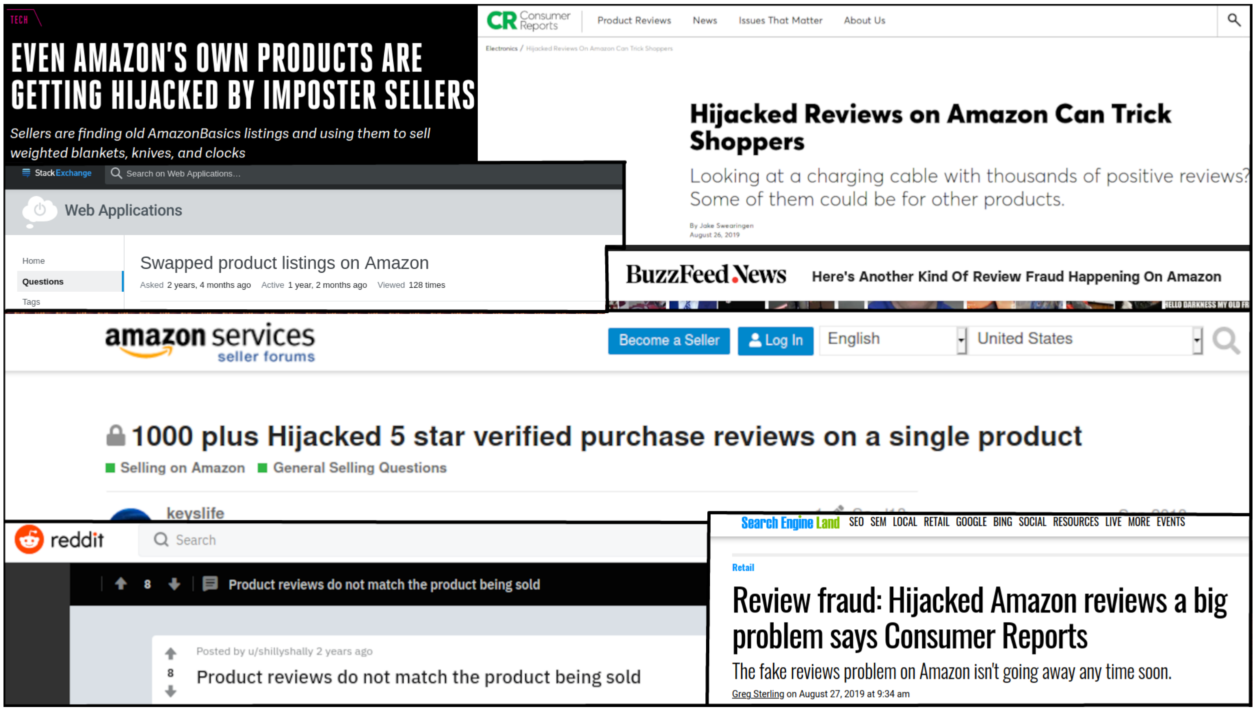
[redacted]
best cat litter ever
Reviewed in the United Kingdom on November 8, 2020
Verified Purchase

Figure 2.2: Example of Global Products Review Hijacking - Lick Mats

Figure 2.2 shows an example hijacking of a product earlier sold in a foreign land. All the reviews from an earlier country (United Kingdom) are for cat litter while the product is lick mat and all US reviews are for that lick mat.

2.2 Reports in News Media and Examples

Review Hijacking has been a highly reported fraud technique on Amazon since 2019 [24]. There have been reports on eBay as well [11]. Many bloggers and tech enthusiasts have spotted this review fraud. News journalist sources and user anecdotes like Consumer Reports [3], Stack Exchange [25], BuzzFeed [2], etc., have widely reported on the existence of this kind of review manipulation. Some tech bloggers call them "bait-and-switch" reviews [26] [27] as customers are baited with many positive reviews but receive inferior products. Consumer Reports has also started a petition to present to Amazon for stopping this nuisance [28]. These reports are also trending on Twitter with #stopreviewhijacking [29]. Even Amazon basics' products are getting hijacked by impostor sellers [30]. Figure 2.3 shows some examples of review news media that have reported review hijacking.



TECH

EVEN AMAZON'S OWN PRODUCTS ARE GETTING HIJACKED BY IMPOSTER SELLERS

Sellers are finding old AmazonBasics listings and using them to sell weighted blankets, knives, and clocks

Electronics / Hijacked Reviews On Amazon Can Trick Shoppers

Hijacked Reviews on Amazon Can Trick Shoppers

Looking at a charging cable with thousands of positive reviews? Some of them could be for other products.

By Jake Swearingen
August 26, 2019

Swapped product listings on Amazon

Asked 2 years, 4 months ago Active 1 year, 2 months ago Viewed 128 times

Here's Another Kind Of Review Fraud Happening On Amazon

1000 plus Hijacked 5 star verified purchase reviews on a single product

Posted by u/shillyshally 2 years ago

Product reviews do not match the product being sold

Retail

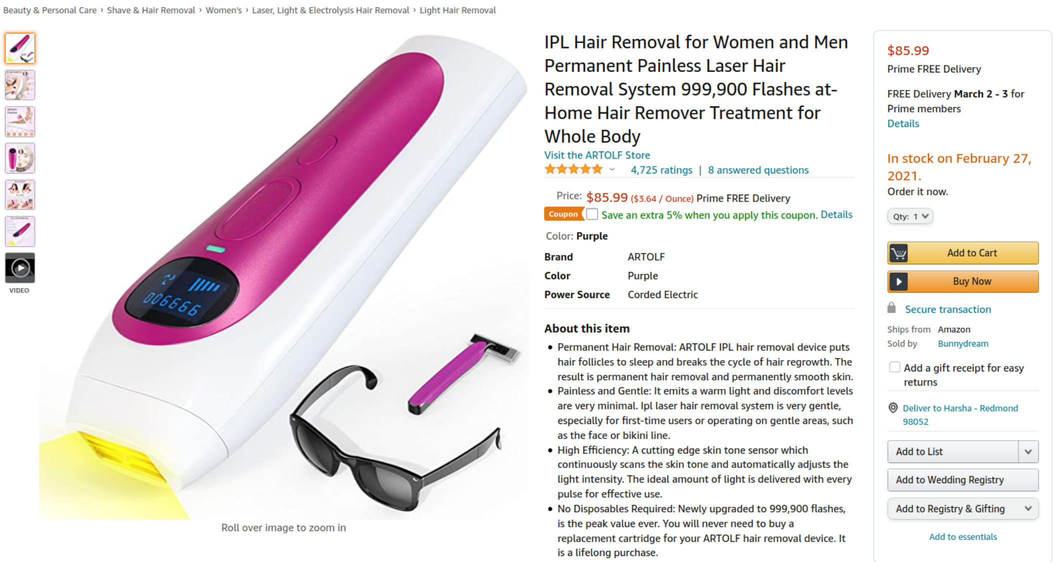
Review fraud: Hijacked Amazon reviews a big problem says Consumer Reports

The fake reviews problem on Amazon isn't going away any time soon.

Greg Sterling on August 27, 2019 at 9:34 am

Figure 2.3: Tech News

Some examples of review hijacking are as follows:



Beauty & Personal Care › Shave & Hair Removal › Women's › Laser, Light & Electrolysis Hair Removal › Light Hair Removal

IPL Hair Removal for Women and Men Permanent Painless Laser Hair Removal System 999,900 Flashes at-Home Hair Remover Treatment for Whole Body

Visit the ARTOLF Store

★★★★★ 4,725 ratings | 8 answered questions

Price: $85.99 ($3.64 / Ounce) Prime FREE Delivery

Coupon □ Save an extra 5% when you apply this coupon. Details

Color: **Purple**

Brand	ARTOLF
Color	Purple
Power Source	Corded Electric

About this item

- Permanent Hair Removal: ARTOLF IPL hair removal device puts hair follicles to sleep and breaks the cycle of hair regrowth. The result is permanent hair removal and permanently smooth skin.
- Painless and Gentle: It emits a warm light and discomfort levels are very minimal. Ipl laser hair removal system is very gentle, especially for first-time users or operating on gentle areas, such as the face or bikini line.
- High Efficiency: A cutting edge skin tone sensor which continuously scans the skin tone and automatically adjusts the light intensity. The ideal amount of light is delivered with every pulse for effective use.
- No Disposables Required: Newly upgraded to 999,900 flashes, is the peak value ever. You will never need to buy a replacement cartridge for your ARTOLF hair removal device. It is a lifelong purchase.

$85.99

Prime FREE Delivery

FREE Delivery **March 2 - 3** for Prime members

Details

In stock on February 27, 2021.

Order it now.

Qty: 1

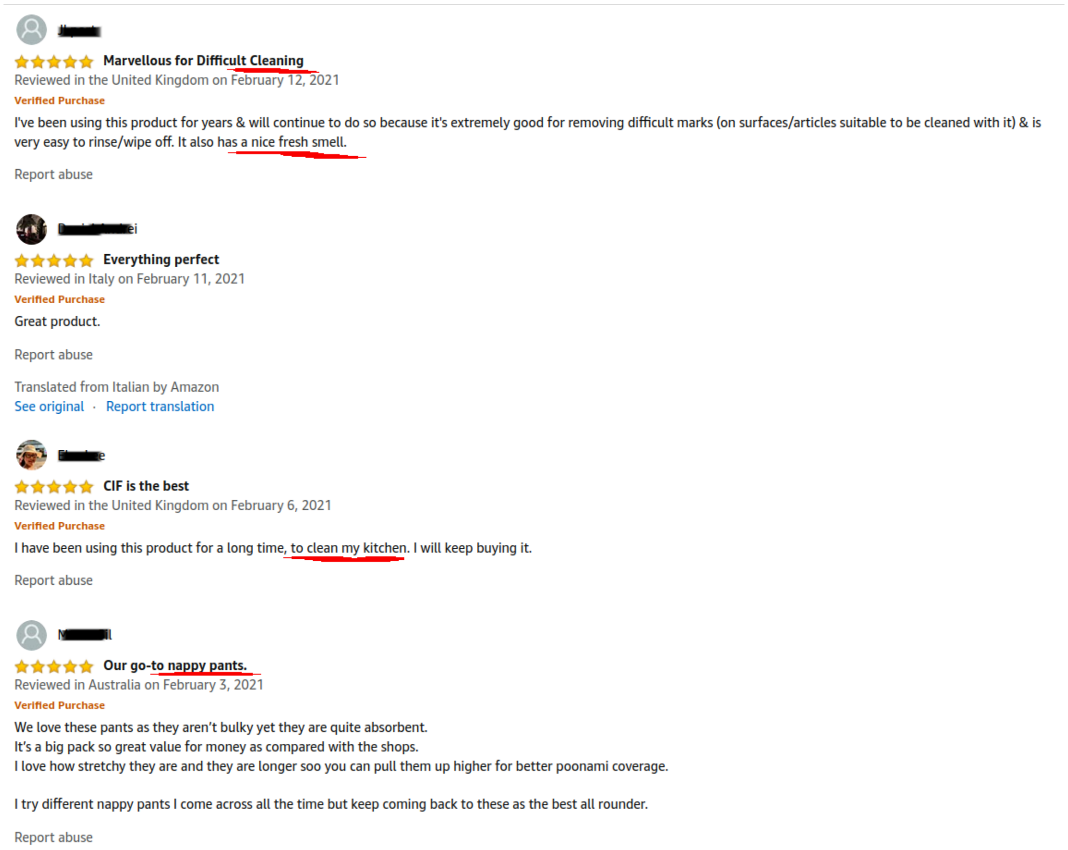
Ships from Amazon

Sold by Bunnydream

□ Add a gift receipt for easy returns

Deliver to Harsha - Redmond 98052

Figure 2.4: Example Product : Hair Removal Machine



[redacted]

★★★★★ **Marvellous for Difficult Cleaning**

Reviewed in the United Kingdom on February 12, 2021

Verified Purchase

I've been using this product for years & will continue to do so because it's extremely good for removing difficult marks (on surfaces/articles suitable to be cleaned with it) & is very easy to rinse/wipe off. It also has a nice fresh smell.

Report abuse

[redacted]

★★★★★ **Everything perfect**

Reviewed in Italy on February 11, 2021

Verified Purchase

Great product.

Report abuse

Translated from Italian by Amazon

See original · Report translation

[redacted]

★★★★★ **CIF is the best**

Reviewed in the United Kingdom on February 6, 2021

Verified Purchase

I have been using this product for a long time, to clean my kitchen. I will keep buying it.

Report abuse

[redacted]

★★★★★ **Our go-to nappy pants.**

Reviewed in Australia on February 3, 2021

Verified Purchase

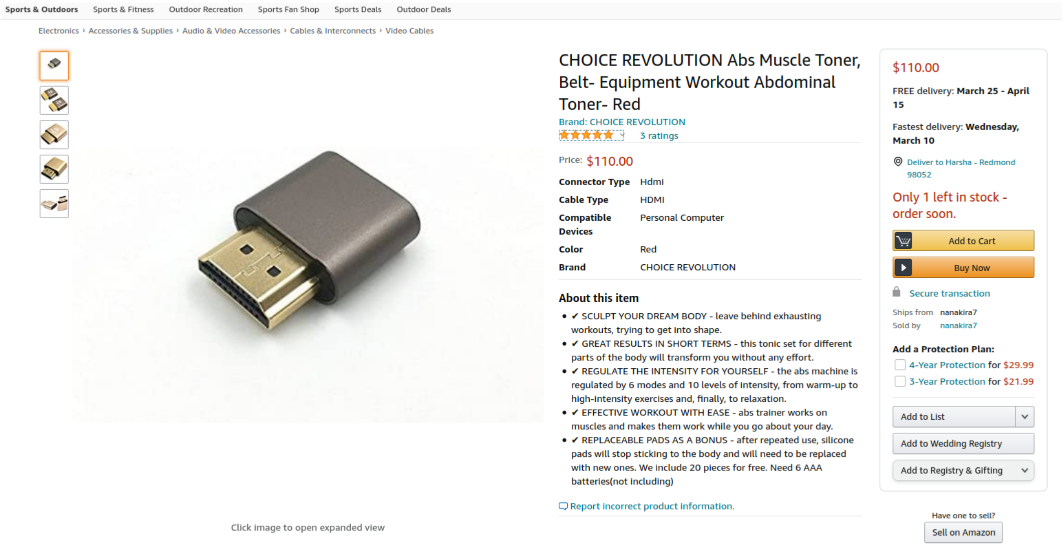
We love these pants as they aren't bulky yet they are quite absorbent.
It's a big pack so great value for money as compared with the shops.
I love how stretchy they are and they are longer soo you can pull them up higher for better poonami coverage.

I try different nappy pants I come across all the time but keep coming back to these as the best all rounder.

Report abuse

Figure 2.5: Example Hijacked Reviews for Hair Removal Product

Figure 2.4 shows an example with a product which is a hair remover. When we read reviews for this product in Figure 2.5, we see reviews of kitchen cleaners and nappy pants merged with those of the hair remover. We can also see that all these reviews are verified purchases.



Sports & Outdoors
Sports & Fitness
Outdoor Recreation
Sports Fan Shop
Sports Deals
Outdoor Deals
Electronics › Accessories & Supplies › Audio & Video Accessories › Cables & Interconnects › Video Cables
Click image to open expanded view

CHOICE REVOLUTION Abs Muscle Toner, Belt- Equipment Workout Abdominal Toner- Red

Brand: CHOICE REVOLUTION

★★★★★ 3 ratings

Price: $110.00

Connector Type	Hdmi
Cable Type	HDMI
Compatible Devices	Personal Computer
Color	Red
Brand	CHOICE REVOLUTION

About this item

- ✓ SCULPT YOUR DREAM BODY - leave behind exhausting workouts, trying to get into shape.
- ✓ GREAT RESULTS IN SHORT TERMS - this tonic set for different parts of the body will transform you without any effort.
- ✓ REGULATE THE INTENSITY FOR YOURSELF - the abs machine is regulated by 6 modes and 10 levels of intensity, from warm-up to high-intensity exercises and, finally, to relaxation.
- ✓ EFFECTIVE WORKOUT WITH EASE - abs trainer works on muscles and makes them work while you go about your day.
- ✓ REPLACEABLE PADS AS A BONUS - after repeated use, silicone pads will stop sticking to the body and will need to be replaced with new ones. We include 20 pieces for free. Need 6 AAA batteries(not including)

Report incorrect product information.

$110.00

FREE delivery: **March 25 - April 15**

Fastest delivery: **Wednesday, March 10**

Deliver to Harsha - Redmond 98052

Only 1 left in stock - order soon.

Ships from nanakira7
Sold by nanakira7

Add a Protection Plan:

□ 4-Year Protection for $29.99
□ 3-Year Protection for $21.99

Figure 2.6: Example of Hijacking Midway - USB / Abs Toner

Here, figure 2.6 shows a product getting hijacked. The product is a gray electronic USB driver in the image but, the title suggests a red abs muscle toner. Next, the seller would change the image to abs toner to hijack it completely.

2.3 Related Work

We do not have any previous related work in this particular domain. Hence, we shall explore some papers for other tasks we can use.

One way of catching review hijacking is by matching a product title and description to its review text. Higgins et al. [34] developed models for an essay rating system to detect bad-faith essays, to compare the essay titles to the essay text to determine whether the title was for the same or not by using similarity of words. Louis et al. [35] worked on whether a particular essay was related to the essay prompt or question. She did it by expanding short prompts and spell correcting the texts. Rei et al. [36] extended this work and combined various sentence similarity measures like TF-IDF and Word2Vec embeddings with moderate improvement over Higgins' work. Ryu et al. [37] proposed a neural sentence embedding method representing sentences in a low-dimensional continuous vector space that emphasizes aspects in-domain and out-of-domain for a given scenario. Another Deep Learning-based approach to detecting off-topic content comes from Fake News Challenge Stage 1 (FNC-1) to determine if the headline of a news article agrees, disagrees, or is unrelated to the text. Hanselowski et al. [38] used the BiLSTM model with Attention for this task.

2.4 Metrics

Finally, we introduce some of the metrics we shall use in evaluating our proposed machine learning methods in upcoming chapters. There are many metrics available to evaluate ML models in different applications.

2.4.1 Terminology

We shall first share some terminologies :

- True Positive (TP): Number of samples that belong to a class and are classified correctly.
- False Positive (FP): Number of samples that do not belong to the class and are classified incorrectly to be belonging to the class.
- True Negative (TN): Number of samples that do not belong to the class and are classified correctly as negatives.
- False Negative (FN): Number of samples that belong to the class but incorrectly classified as belonging to the class.
- True Positive Rate (TPR): True Positive Rate (TPR, also called sensitivity) is calculated as TP/TP+FN. TPR is the probability that an actual positive will test positive.
- False Positive Rate (FPR): False Positive Rate is a probability of falsely rejecting the null hypothesis for a particular test.

At times, one needs to use more than one metric to evaluate our model correctly. We discuss metrics that we use in section 6 to evaluate our models.

2.4.2 Accuracy

The most common metric for any classification task is the classification accuracy, obtained by dividing the number of correct predictions by the total number of predictions. It is shown mathematically as below.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} = \frac{True\ Positive + True\ Negative}{Total\ Predictions}$$

Researchers use this method widely because it is intuitive, easy to calculate, easy to interpret, and a simple metric to compare. But, this intuition breaks when we have highly skewed data.

As discussed in section 1.4, our model has very few positive values (< 0.01%). If a machine learning model predicts 0 cases of review hijacking when we run an algorithm, our accuracy is more than 99.99%. While superficially, this appears to be an excellent accuracy, the method does not solve the problem at all.

Once we synthetically generate a balanced dataset as discussed in section 5.2, we use classification accuracy as a measure for evaluating supervised learning algorithms.

2.4.3 Precision-Recall

Precision is a class-specific metric that is better for imbalanced classes than accuracy. It is defined as follows:

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

Recall is another important metric that tells us the fraction of cases correctly identified by the ML model.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

As we do not have the labels, we use precision for evaluating the unsupervised algorithms. Also, We use the precision-recall curve (a plot of precision vs. recall) to understand the results from the supervised learning models.

2.4.4 ROC

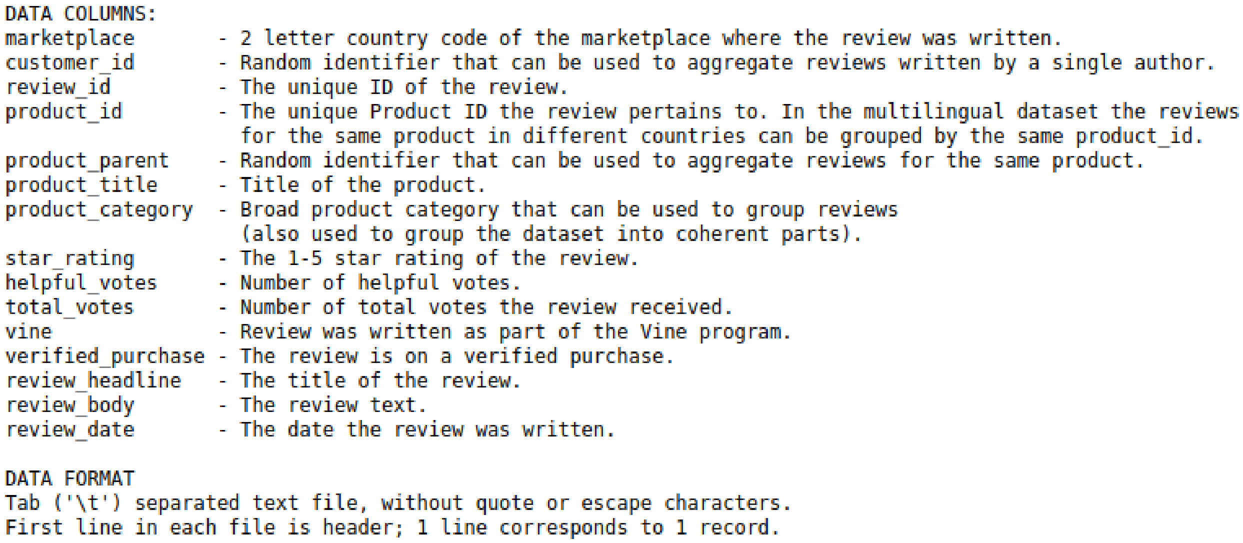
The number of positive samples is minuscule ($< 0.01\%$), and our distribution is highly skewed. Therefore, we use AUC (Area Under Curve) value under the ROC curve for the results. We obtain the Receiver Operator Characteristic (ROC) curve by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). It balances the trade-off between sensitivity (or TPR) and specificity (1 - FPR) and hence is a commonly used metric for skewed datasets.

3. DATASETS

The first thing we need for any research problem is a sufficient and related dataset. There are several different types of datasets available for Amazon reviews from various sources like Kaggle, data.world, and GitHub. We required a dataset that consisted of Amazon product titles and review texts at the very least. We primarily used two datasets for our experiments: (i) a public dataset shared by Amazon; and (ii) an Amazon dataset crawled by Julian McAuley at the University of California, San Diego.

3.1 Dataset-1: Amazon Customer Reviews Public Dataset by Amazon

Amazon has released a dataset for 130+ million reviews written on the Amazon.com marketplace from 1995 to 2015 [39]. The data is available in tab-separated TSV files and column-oriented parquet format in the amazon-reviews-pds S3 bucket in AWS (Amazon Web Services). They have provided the data separated into multiple TSV files containing different categories of products like electronics, books, groceries, etc.



```
DATA COLUMNS:
marketplace       - 2 letter country code of the marketplace where the review was written.
customer_id       - Random identifier that can be used to aggregate reviews written by a single author.
review_id         - The unique ID of the review.
product_id        - The unique Product ID the review pertains to. In the multilingual dataset the reviews
                    for the same product in different countries can be grouped by the same product_id.
product_parent    - Random identifier that can be used to aggregate reviews for the same product.
product_title     - Title of the product.
product_category  - Broad product category that can be used to group reviews
                    (also used to group the dataset into coherent parts).
star_rating       - The 1-5 star rating of the review.
helpful_votes     - Number of helpful votes.
total_votes       - Number of total votes the review received.
vine              - Review was written as part of the Vine program.
verified_purchase - The review is on a verified purchase.
review_headline   - The title of the review.
review_body       - The review text.
review_date       - The date the review was written.

DATA FORMAT
Tab ('\t') separated text file, without quote or escape characters.
First line in each file is header; 1 line corresponds to 1 record.
```

Figure 3.1: Amazon Customer Reviews Public Dataset

We use this dataset for our unsupervised learning methods Inverse Boolean retrieval, Topic modeling, TF-IDF, and BM25 (Chapter 4). We started by using this dataset due to its ease of access and use. This dataset was the first one we found, was indexed well, and officially shared by Amazon.com, and hence we started working on it. Being in TSV format, it is easier to read and write using pandas without extra overhead. The TSV files contained product titles, so we did not have to worry about merging datasets. The reviews were also separated category-wise and hence easier to work on one category at a time. They also provided a small sample file to get started.

The column names on this dataset are: marketplace, customer_id, review_id, product_id, product_parent, product_title, product_category, star_rating, helpful_votes, total_votes, vine, verified_purchase, review_headline, review_body and review_date as seen from figure 3.1. We only used the reviews from the US marketplace and used the following columns:

- review_id: A unique identifier (ID) of the review. We used this optionally when it is required to identify a particular review. Generally, indexes worked just fine.

- product_id: The ID of the product for which the review is written, also known as ASIN. An Amazon Standard Identification Number is a 10-character alphanumeric unique identifier assigned by Amazon.com. It is used for identifying products within Amazon [40] [41].

- product_title: The title of the product provided by the seller. This is a major field as we used this text string as a query (Q) for our IR unsupervised methods. More details in chapter 4.

- star_rating: The star rating given by the customer. It is an integer with values 1 to 5. We used this to optionally discard products with lower ratings, as no seller would prefer hijacking a listing with lower ratings.

- review_headline: Headline/title of the review. This is merged with review_body to form review_text.

- review_body: The actual text of the review provided by the customer. This is used as document (D) for our IR unsupervised methods. More details in chapter 4.

3.2 Dataset-2: Amazon Review Data by J. Ni and J. McAuley

This dataset contains 233.1 million reviews from May 1996 to October 2018 [42]. It is an updated version of a previous dataset provided by the same group in 2014. The dataset consists of two JSON files, one contains the reviews, and the other contains metadata of the products.

We used this dataset majorly due to the availability of newer review data. The first report of Review Hijacking was in 2018. Intuitively sellers may have only started using this technique recently. Presumably, reviews from 2018 should result in a higher number of hijacked reviews. Also, newer review listings will provide more insight into the present-day review system. Another reason for selecting this dataset was the presence of product descriptions. Reviews are essentially an analysis provided by the customers. Thus, now we could compare these two texts to obtain a similarity score and determine review hijacking. The product descriptions helped using the dataset for our supervised learning methods Siamese LSTM network and BERT sequence pair classifier (Chapter 5) There are 100 million more reviews in this dataset than in the previous dataset. Hence, we can now work with more data as well.

The disadvantage here is that the data is in JSON format that requires loading, formatting, and pre-processing to be done for storing in pandas dataframes. Again, this data is separated in two files, product metadata and review data. These big files also caused RAM issues.

The review data is available in multiple forms; one complete raw file, a rating-only version, a 5-core subset containing users and items having at least five reviews (75.26 million reviews), and per category separated data. We used the super-set raw file, 5-core subset, and per-category files as and when needed.

The review data consists of the columns : 'reviewerID', 'asin', 'reviewerName', 'vote', 'style', 'reviewText', 'overall', 'summary', 'unixReviewTime', 'reviewTime', 'verified', 'image'. We use only the following columns for our methods :

- 'asin': ASIN of the product. We use this as an identifier to merge with the product metadata.
- 'style': Optional product features like color, size, etc. We append this with reviewText if

present. This field is present for very few products and does not impact results much.

- ’reviewText’: Text in the review body
- ’summary’: Summary or headline of the review. We concatenate this at the beginning of the reviewText.

The metadata dataset contains 15.5 million products and their features such as descriptions, category information, price, brand, etc.

The columns in the metadata file are: ‘category’, ‘description’, ‘title’, ‘image’, ‘brand’, ‘feature’, ‘rank’, ‘main_cat’, ‘asin’, ‘also_buy’, ‘also_view’, ‘similar_item’, ‘price’, ‘date’, ‘details’, ‘tech1’, ‘tech2’, ‘fit’. We used only the following columns for our methods:

- ‘description’: Product description provided by the seller. It forms a considerable part of ’product_text’.
- ‘title’: Title or name of the product. It is concatenated at the beginning of the ‘product_text’.
- ‘brand’: Brand of the product. Concatenated with ‘product_text’
- ‘feature’: These are product features like color, size, etc., and are provided for very few products. We concatenated with the ‘product_text’ if present, but they did not provide much difference in results.
- ‘asin’: ASIN of the product.

3.3 Data Analysis

We here perform data analysis of dataset-2. We first begin with some rudimentary analysis of the dataset. We discuss histograms of the number of products in each category, the number of reviews per product, and the average number of reviews per product.

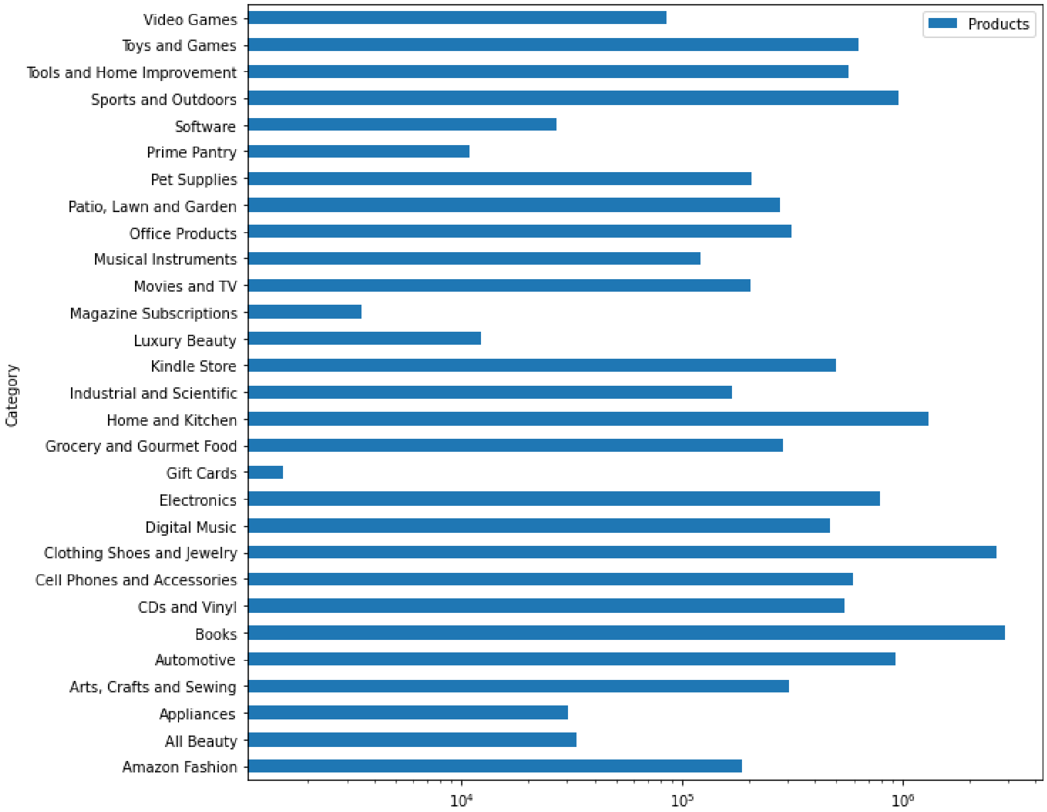


Figure 3.2: Products per Category (log scale)

Figure 3.2 represents the number of products per category on a logarithmic scale. We see that number of products is quite varied. The number of “Books”, “Clothing shoes and jewelry” and “Home and Kitchen” are highest (> 1M) while Gift cards and magazine subscriptions are lowest (< 10K). Generally, every category has products in the range of 10K to 1M.

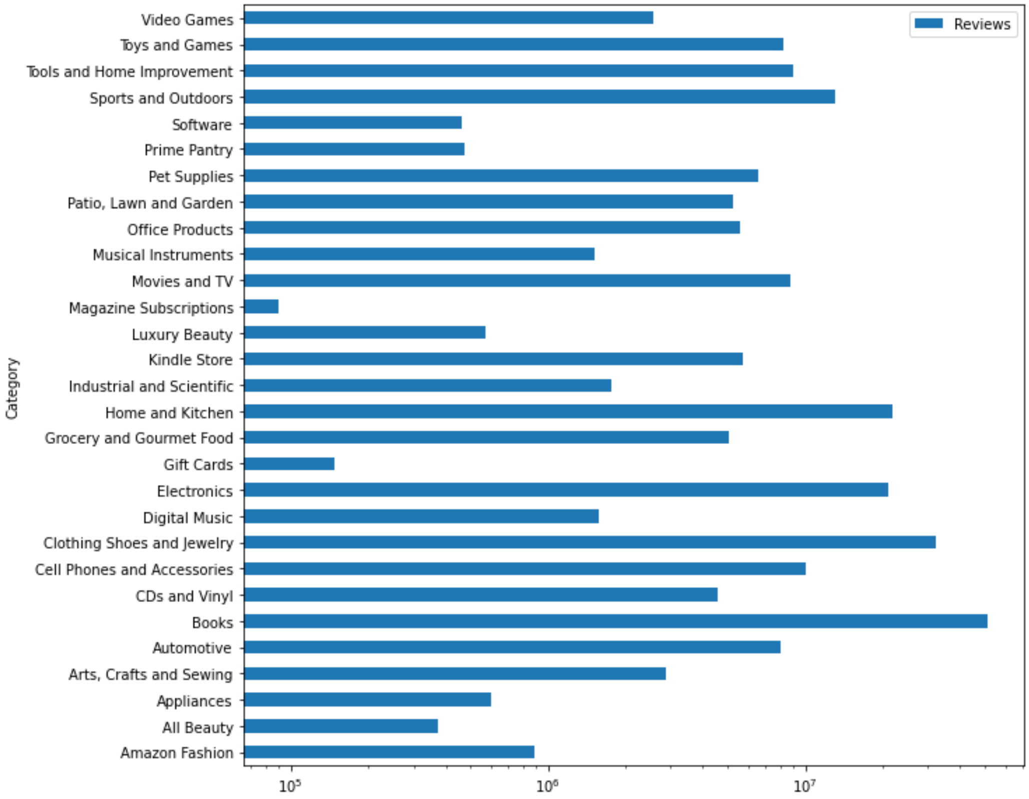


Figure 3.3: Number of Reviews per Product (log scale)

Figure 3.3 shows the distribution of reviews per product. The distribution looks similar to figure 3.2. This similarity indicates that reviews per product category are not very varied.

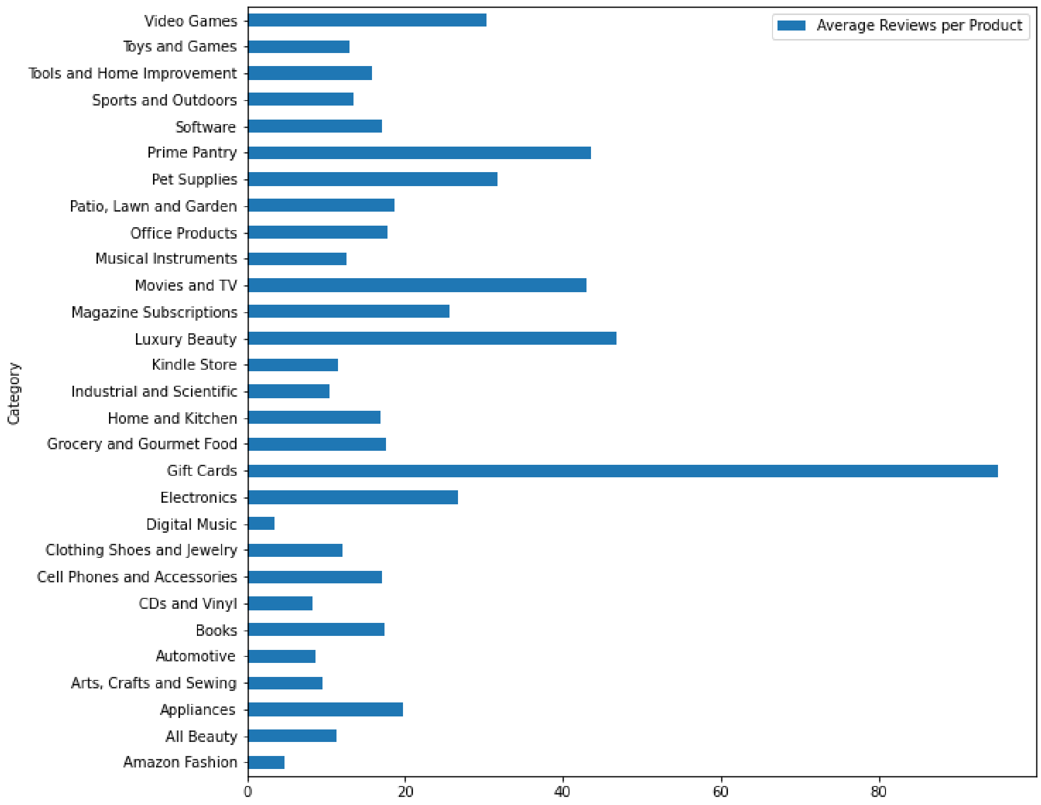


Figure 3.4: Average Reviews per Product

Figure 3.4 shows the average reviews per product. We see that these values are in the range of 10^5 to 10^8. We observe that gift cards have the highest average reviews per product. This high number might also be because Amazon has a limited number of gift cards, so the number of reviews is higher per product. Other high reviewed categories are the prime pantry, Movies & TV and Luxury beauty, where products are less in number. Products that are less in number tend to be reviewed more, so perhaps even bought more, as there are limited options.

We will now want to see what kind of words are present in our reviews and product descriptions to check what kind of approach we can take for our machine learning methods.

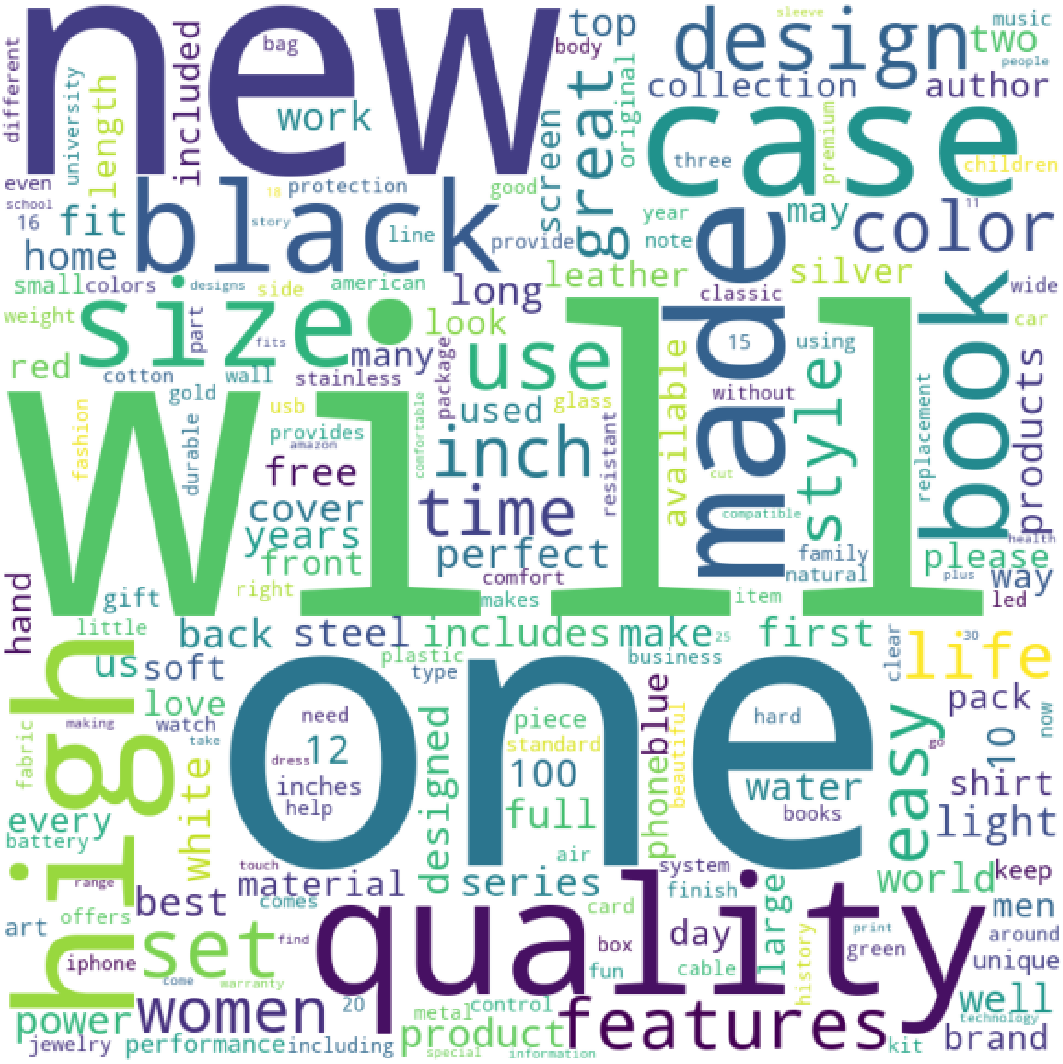


Figure 3.5: Word Cloud of all Product Titles and Descriptions

The word cloud in figure 3.5 shows that the most common words in the product title and description (combined) are ‘will’, which shows that the sellers generally discuss future tense as when shoppers will buy their products. All other most common words are describing words like ‘new’, ‘one’, ‘high quality’, ‘black’, ‘size’, etc. Other words are products themselves like ‘case’, ‘book’, etc. We see the word ‘women’, which shows that a good portion of products must be targeting women.

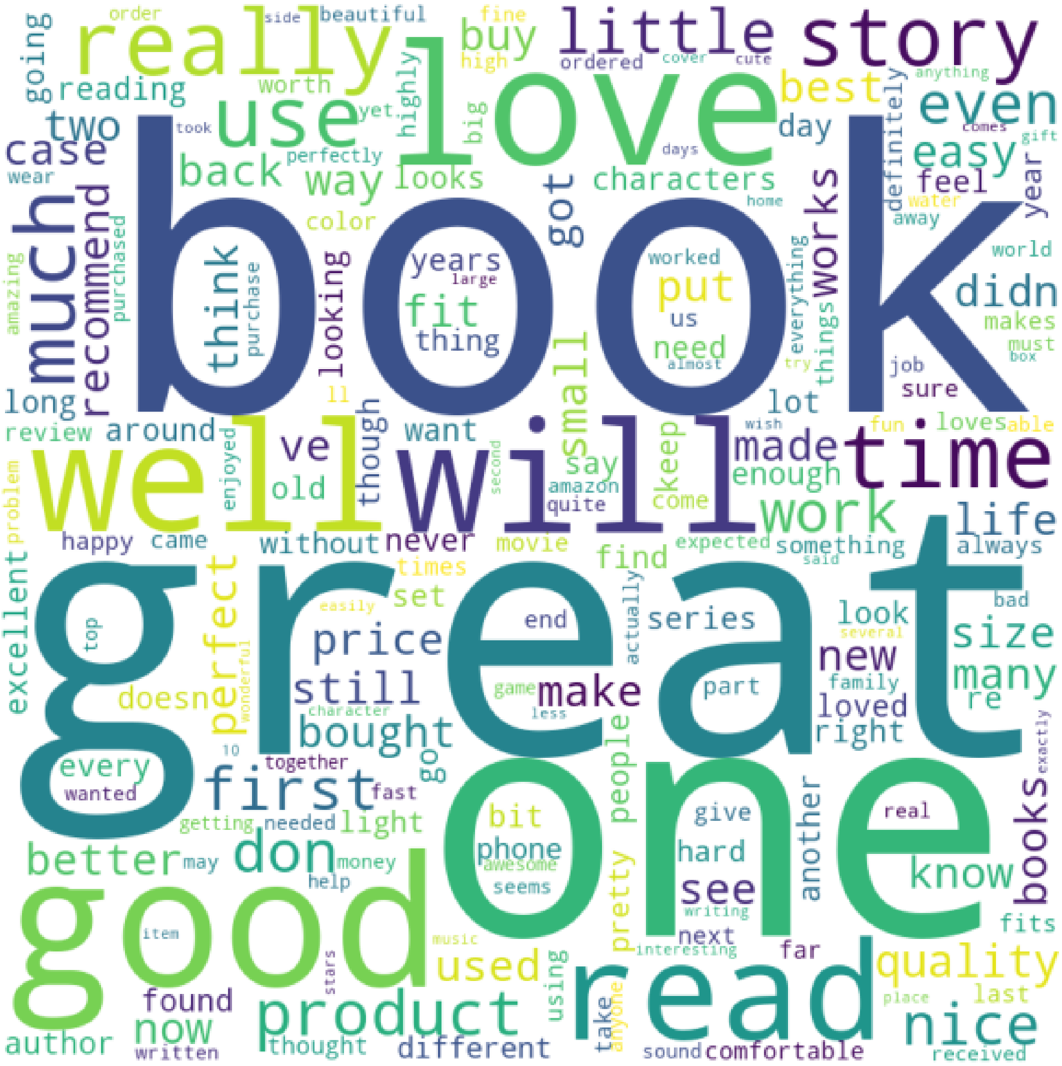


Figure 3.6: Word Cloud of all Review Text

The word cloud in figure 3.6 shows the most common words in the reviews. As there are major reviews in the book category, ‘book’ is the most common word. Thereafter, the most common words are ‘great’, ‘one’, ‘good’, ‘will’, ‘well’, ‘love’ and so on. We see a majority of the words are descriptive words. This observation provides a major analysis that reviews are essentially descriptions of the products. Hence, we can use similarities between product descriptions and reviews for our learning algorithms.

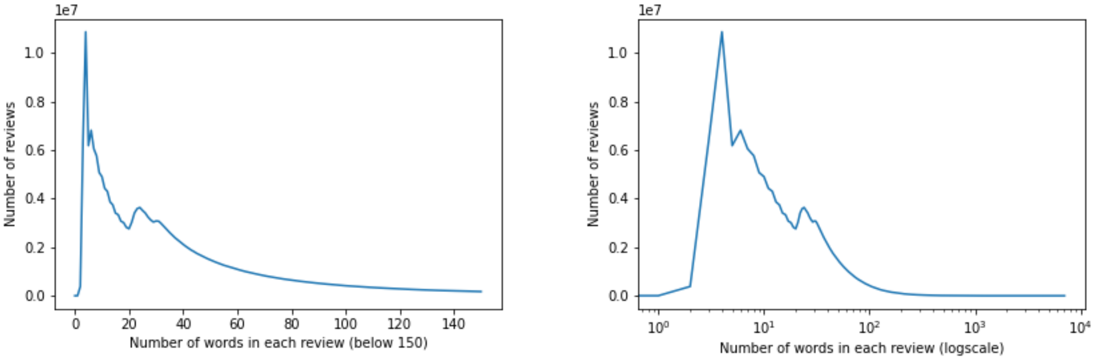


Figure 3.7: Number of Words in Summary + Review Text

Figures in 3.7 show that the graph for the number of words in the review is highly skewed. The first graph shows only reviews with the number of words below 150 words, and the second one shows the same graph in log-scale. The number of words in reviews ranges from 0 to ≈7020 words. Hence, when data cleaning should delete some reviews with 0 words. The figures show that the texts are generally short, about 1 to 10 words. 75 %ile values in range 0 to 60. These statistics show that people write short reviews for the products in general. There is also a second spike between 20-40 words, which is probably the section where people write good, brief, and analyzing reviews. The mean of the number of words is ≈55 words, and the median is 29 words. There are a few people who would write thousands of words for reviews.

Similarly, figures in 3.8 show that the majority of reviews have their length in the range of 10 to 400. 75%ile values are below 318. These graphs are also skewed but less than that of the number of words. Even if the number of words is not high, the text length is sufficiently large. Our methods will have enough review text to predict hijacking or otherwise.

Figures in 3.9 are box plots for the number of words and length of review text. We can see that number of words is more skewed than the length of words.

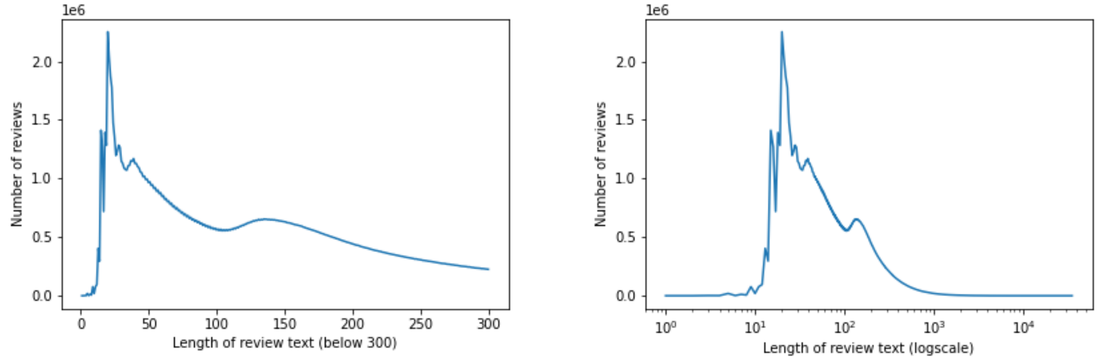


Figure 3.8: Length of Summary + Review Text

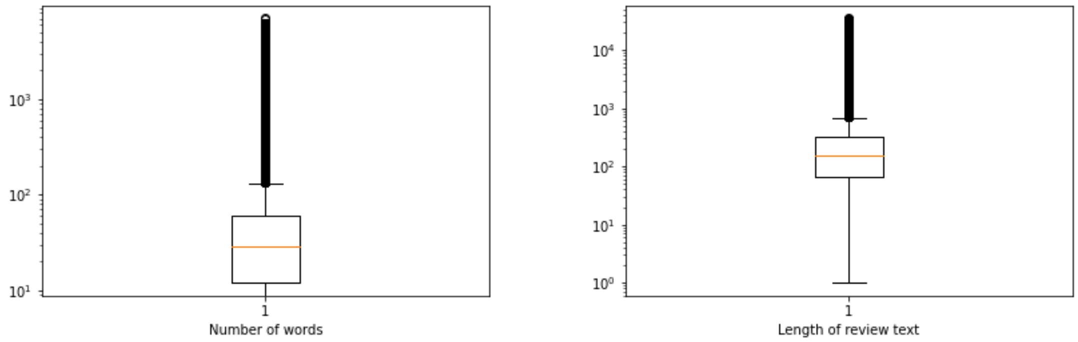


Figure 3.9: Box Plots for Number of Words in Review Text and Length of Review Text

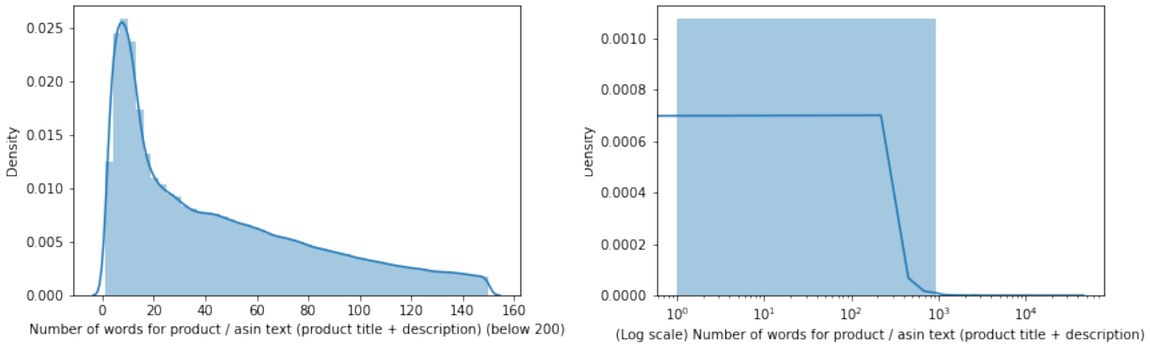


Figure 3.10: Number of Words in Product Title + Description

Figures in 3.10 show that the graph for the number of words in product title and description are highly skewed like in review text. The first graph shows only reviews below 150 words, and the second one shows the same graph in log-scale. The number of words in reviews range from 1 to 46230 words. This distribution shows that there are a bunch of sellers who write huge reviews for more than 10K words. 75 %ile values are below 100, and 99%ile values are below 500. The mean of the number of words is 82 words, and the median is 46 words. These statistics show that sellers generally write a descriptive description to attract more customers. There is a single peak at about 1 to 20 words, which indicates that many products either have very short or no description at all (only title) in our dataset.

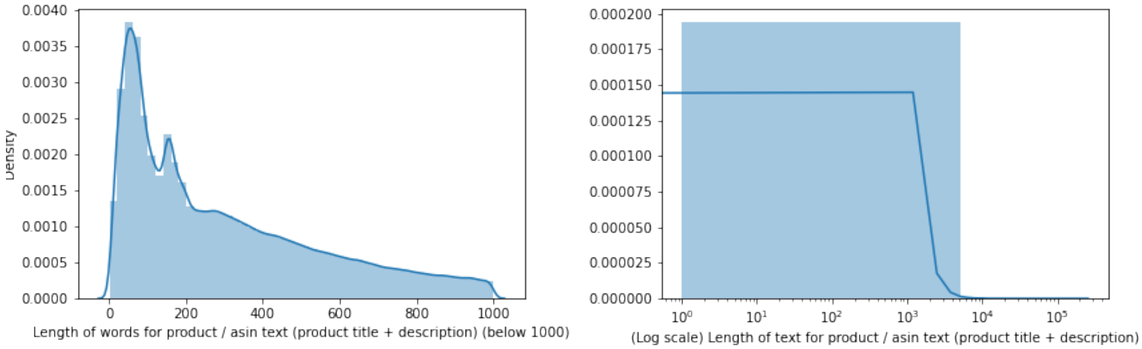


Figure 3.11: Length of Summary + Review Text

Figures in 3.11 show that along with the number of words, the length of the product description is also very skewed. The first graph shows the lengths of product text below 1000, and the second graph shows in log-scale. The product text length ranges from 1 to 256557. The mean of the data is 517, and the median is 288 characters. 75 %ile values are below 650, and 99%ile values are below 3250. Hence, a small percentage of sellers write huge reviews of more than 100k in length. Generally, they write some 100s of words for product description. We observe two peaks, first between 1 to 100 and the second between 150 to 200 text length.

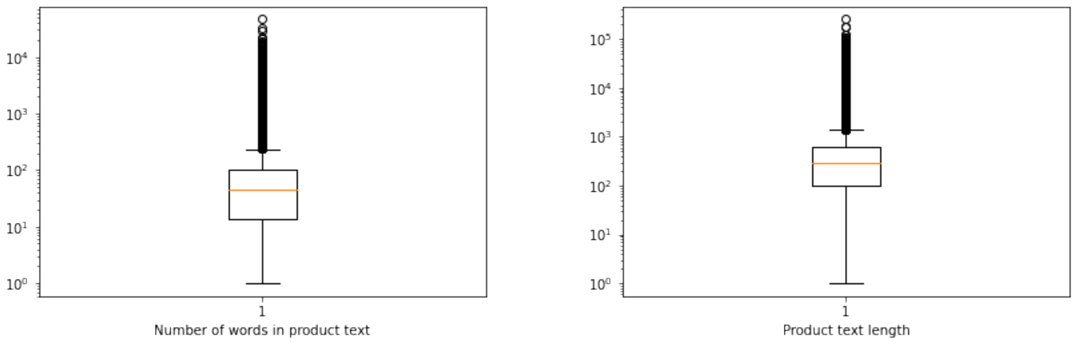


Figure 3.12: Box plots for Number of Words in Product Text and Length of Product Text

Figures in 3.12 are box plots for the number of words and length of product text. We can see that both numbers of words and product text are highly skewed. There are a lot of outlier points as well.

3.4 Summary

Table 3.1 presents the major difference between the datasets. Both have their advantages and disadvantages.

Dataset-1	Dataset-2
Total 130+ million reviews	Total 233.1 million reviews
from 1995 to 2015 (older reviews)	from May 1996 - Oct 2018 (newer reviews)
Shared by Amazon.com	Shared by UCSD group
Is in TSV or parquet format. Hence, easier to access	Is in json format, need pre-processing to read.
Does not contain product description	Contains product description
Reviews available category wise	Reviews available category-wise and as a bulk raw form
Better formatted	Formatting not as good as dataset-1

Table 3.1: Dataset Comparison Summary

We choose dataset-1 for unsupervised machine learning methods and then dataset-2 for supervised learning methods. We did not need product descriptions for unsupervised learning algorithms, so we used the simple dataset. We needed product descriptions for supervised ones, so we used dataset-2.

4. UNSUPERVISED INVESTIGATION

We have datasets that contain products and their associated reviews, but we do not know which hijacked review listings in it and which are not. We are unsure if there are such off-topic listings in the given datasets. Hence, we begin our investigation by exploring several unsupervised learning approaches.

Figure 4.1 shows us the plan we use for our unsupervised investigation. We would first perform data cleaning and then apply the following unsupervised methods on this cleaned data:

1. (Inverse) Boolean model of Information Retrieval (BIR)
2. TF-IDF and BM25
3. Named Entity Recognition (NER) with BIR
4. Topic modeling

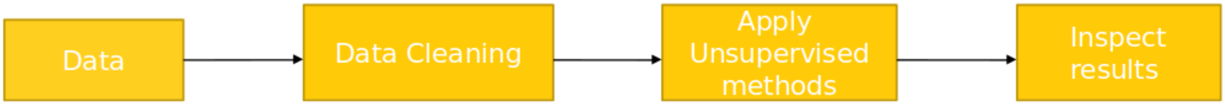


Figure 4.1: Plan for Unsupervised Investigation

It is worth noting that the first three methods here are Information Retrieval (IR) methods, and the fourth, Topic Modeling, is a text summarization method. We will inspect results obtained from these methods to uncover evidence and the extent of review hijacking in our dataset.

4.1 Data Cleaning

As discussed in chapter 3, we used dataset-1 shared by Amazon and performed data cleaning. We dropped unnecessary columns like helpful_votes, verified_purchase, etc., and used only the columns shared in section 3.1. We removed rows with no review text. We removed products with fewer than five reviews as such products are not very useful for our methods. We removed products with overall ratings below 3.5 stars as no seller would want to hijack listings with bad reviews, as this might not be useful for them.

4.2 Information Retrieval Methods

Information retrieval (IR) constitutes methods to search the information resource (a document or several documents) for the relevant information. These methods are used in contexts like web searches. A query (Q), which represents our information need, is matched with the documents (D) in the data resource. As seen in figure 4.2, the IR system takes user query and index documents as input. We score each document on how related the query is to that document. Once we get these scores, we rank the documents in the decreasing order of score values (or increasing relevance). Then, our information retrieval method returns the results with documents of maximum scores.

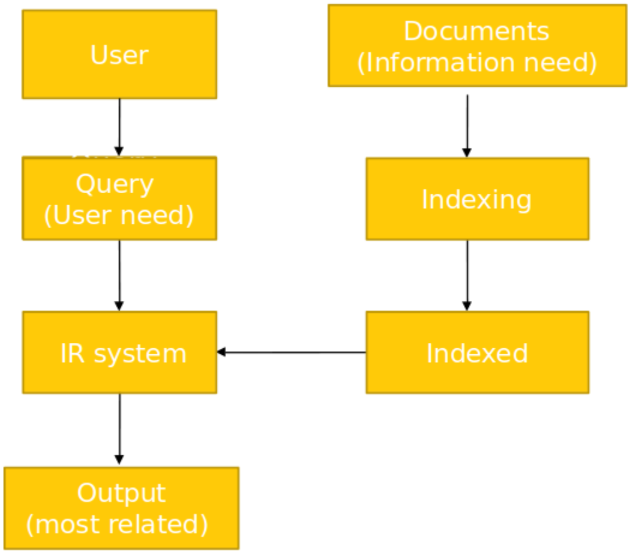


Figure 4.2: Information Retrieval (IR) Method

We use similar techniques, as shown in figure 4.3, here for identifying potential examples of review hijacking. We have no user. We use product title as query Q and review texts as documents

D. Similarly, we match all documents (reviews attached to the product) and score them. But, here, we make a tweak in ranking. We rank the results in increasing order of score values (decreasing order of relevance) and return the results with minimum scores. This ranking shows the results with reviews least matching the product title. If a product contains many feedbacks with low relevance, we hypothesize that this is a hijacked review listing from some other product.

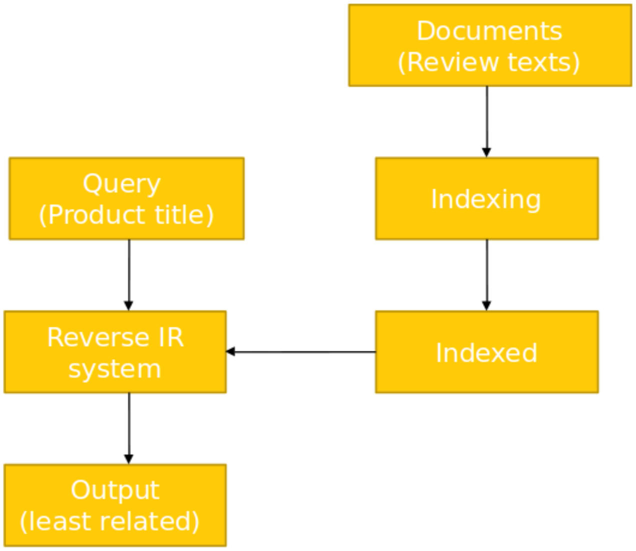


Figure 4.3: Tweaked Reverse IR Technique

We start by implementing simple techniques like the Boolean model of Information Retrieval (BIR) [43], TF-IDF (term frequency-inverse document frequency), and Okapi BM25.

4.2.1 (Inverse) Boolean model of Information Retrieval (BIR)

The first approach is the Boolean model of Information Retrieval, also known as Boolean retrieval [43]. It is based on boolean logic and set theory. In this method, both the query (Q) and the document (D) are considered sets or bags-of-words, where the order of terms does not matter. A document is relevant to the query if it contains all the words in the query. If none of the documents have all the words in the query, it ranks according to the number of query terms matching the document.

For example, if one searches for “Kodak Digital Camera”, boolean retrieval returns all the documents satisfying the set operation: “Kodak” AND “Digital” AND “Camera”, i.e. documents containing “Kodak”, “Digital” and “Camera” as the top results. It searches all the documents for all words in the keyword and returns all such documents.

For our approach, we use words in the product title as the keywords of the query (Q) and the review text associated with the product as documents (D). We run set logic and return results in which none of the keywords in the product title (query) match the review text (document). For example, if the product is “Kodak Digital Camera”, we provide results with boolean logic: NOT “Kodak” AND NOT “Digital” AND NOT “Camera”. Thus, we will get reviews that mention neither of the three query keywords.

Approach

First, we concatenated the review summary and review text for each column. The method uses the set-logic or bag-of-words method. Hence, the order of these columns does not matter. Then, we applied data cleaning to this obtained review text by putting it to lower case, removing stop-words, tokenizing, and stemming. Then we formed an inverted index, with the ASINs as the keys and number of common words between the product title and review text as values. Then, we listed the products with no common words, one common word, and two common words to find potential examples of review hijacking. The intuition here is that hijacked review listings will not contain any words common with the product title.

Observations

Through manual inspection, we found some actual samples in the dataset. There were 17% hijacked listings in the output, i.e., we had 17% precision. (We discuss precision in section 2.4). We found several examples of products from a different category altogether.

The output consisted of a majority of false positives due to many reasons. For example, some reviews discussed product properties like “Sound is good” for a music player or “It fits perfectly” for a garment. In this case, we did not have common words, but reviews belonged to the product.

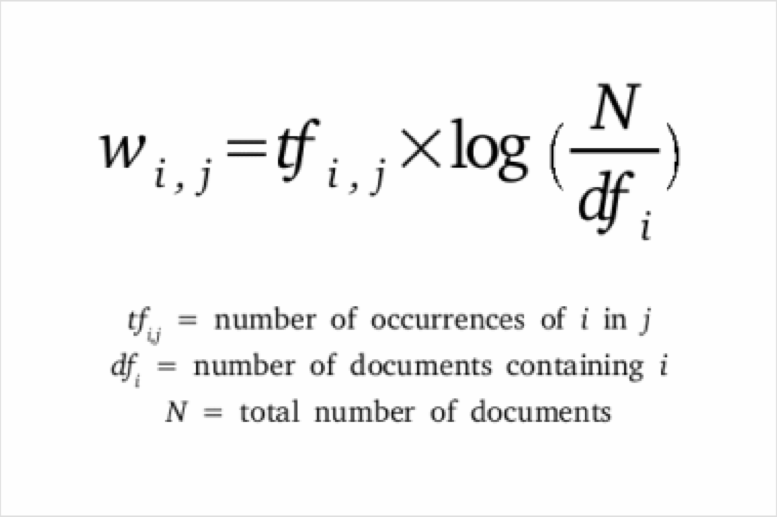
Then, some product reviews were generic, “Great product! Five stars!” or “The product shipped fast”. There is no definite boundary if the reviews belonged to the product or not. A few of the false positives were typing errors.

Now we also searched with some false negatives. We filtered them by one or two common words. Here, we observed that the common words are generally descriptive or stop words. For example, for the product “New 8 GB MP3 Player, With USB Jack and Headphones”, the reviews like “This is a great screen protector. My phone screen looks new.” with the repeated word “new” which is a false negative. Hence, we decided to remove common descriptive words to get better results.

4.2.2 TF-IDF and BM25

Like boolean retrieval, TF-IDF (term frequency-inverse document frequency) and Okapi BM25 (Best Matching 25) score documents based on relevance to the query term and rank them accordingly. These are also bag-of-words models where order does not matter. The prime difference is that the rarer terms have more weight than frequent terms when scoring.

The TF-IDF score increases proportionally the frequency of a word in a document and is offset by the number of documents that contain the word in the corpus. (Refer figure 4.4)



$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Figure 4.4: TF-IDF Formula

In figure 3.6 in chapter 3, we saw that certain words repeat several times in the reviews. The words like “great”, “one”, “good”, and “love” repeat often in reviews as people express their opinions. Product titles may contain these opinionated or describing words in different forms like “new” and “black”. These words might very well be in reviews but do not describe the products enough to determine that review is not hijacked. So we would want these words to weigh less than other words. TF-IDF and BM25 should help with this.

Approach

Similar to the boolean method, we concatenated the review summary and review text for each column and applied data cleaning to this obtained review text. We select the product titles as queries (Q) and review texts attached to them as documents(D). Then we applied TF-IDF for each row and obtained the scores. We performed the same steps for BM25 as well. Then we sorted them in increasing order and considered products with a score less than 2.0 as results.

Observations

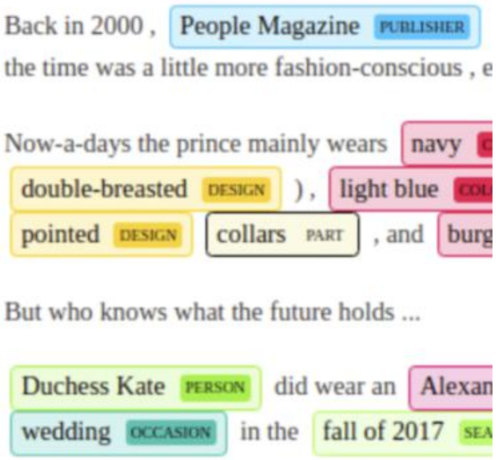
TF-IDF and BM25 gave similar results. They provided better precision than BIR (23%) but lower number of results. The final result still missed quite a few cases while giving many false positives. The recall was better for TF-IDF. (Please refer section 2.4 for precision-recall definition)

The advantage of these methods is that frequent words in product title and review such as “great”, “quality”, and so on, are given a low score, which helps to detect the hijacked reviews better. Changing the TF-IDF score also changes the number of outputs in the final result.

There are a lot of false positives. If there is a frequent product such as “iPhone case”, it treats these as frequent words, hence increasing false negatives. Thus, we see that TF-IDF rates even important words as low score and this method is not sufficient.

4.2.3 Named Entity Recognition (NER) with BIR

We explored Named Entity Recognition (NER) with the boolean retrieval (BIR) approach. A Named Entity is an object in the real world. Named Entity Recognition (NER) is the process of recognizing these entities in structured or unstructured text into predefined categories like names, locations, or organizations. Figure 4.5 shows an example.



Back in 2000 , People Magazine PUBLISHER
the time was a little more fashion-conscious , e

Now-a-days the prince mainly wears navy C
double-breasted DESIGN) , light blue COL
pointed DESIGN collars PART , and burg

But who knows what the future holds ...

Duchess Kate PERSON did wear an Alexan
wedding OCCASION in the fall of 2017 SEA

Figure 4.5: Named Entity Recognition

As we saw earlier, the main disadvantage of boolean retrieval was that adjectives, like, “great” and “good” were getting matched if present in the product title. Hence, we plan to use only nouns in product titles and review text and match only those for better results.

Approach:

The method is the same as the boolean approach, where we merged all reviews and did data cleaning. When preparing the inverted index, we included only nouns and noun phrases instead of all words.

Observations

We obtained only 12% precision (Refer section 2.4). We hoped that no adjectives matched (unlike only boolean retrieval) and all nouns get matched (unlike TF-IDF and BM25). This method still has many false negatives as TextBlob is not very accurate in detecting noun phrases. The limitation is due to the inefficiency of the library.

Again, the product titles are not complete sentences, so detecting noun phrases from this non-sentence structure is not an easy task. Some customers will also write reviews like "I LOVE this bag!!", where the library assumes "LOVE" as a noun because it is in capitals. Nevertheless, the library erred with general sentences as well, and this method was not very useful.

4.3 Topic Modeling

So far, we have been using Information Retrieval (IR) methods and querying product titles to search review text as documents. We now step out of information retrieval methods and take a different approach as an experiment. We attempt to predict the product title from the review text.

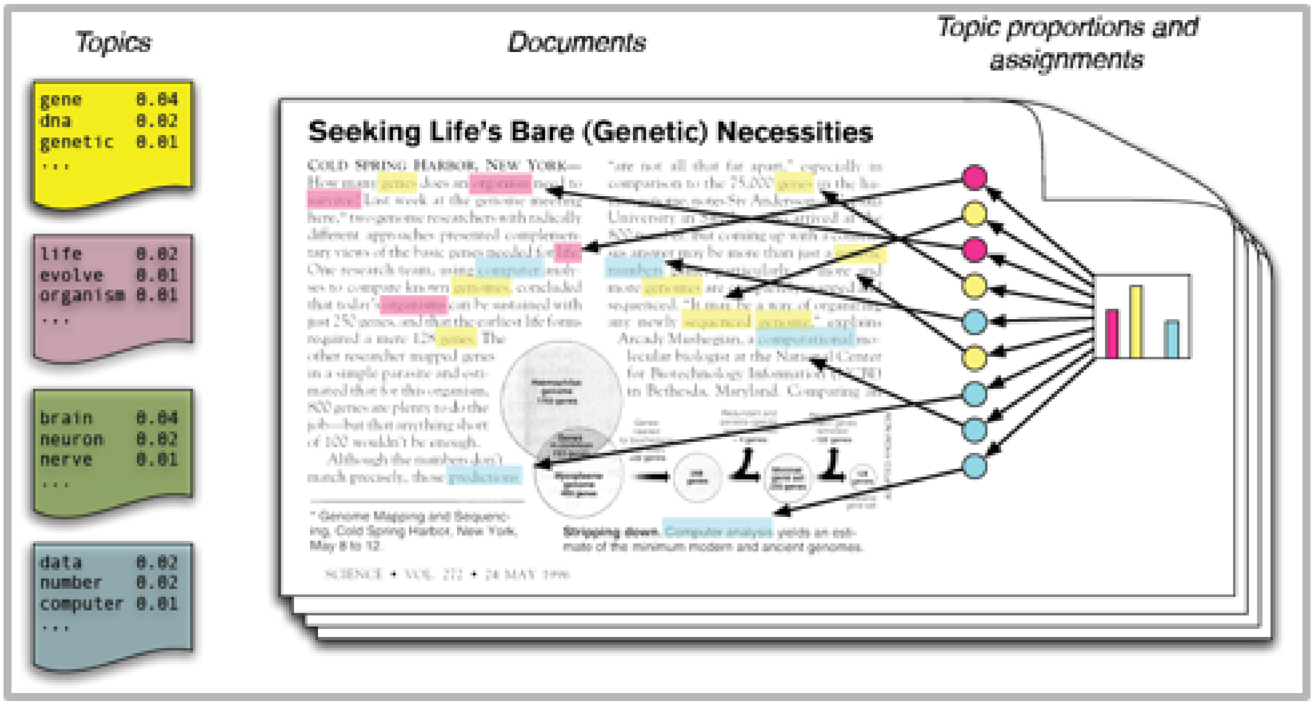


Figure 4.6: Topic Modeling

Topic modeling is a machine learning technique that discovers the “topic” of a paragraph or document based on statistics of words in that document. The intuition behind this text-mining technique is that words like “canine”, “bones”, and “dog” will be more common in a paragraph about dogs. “Dog” can be thus the topic of this paragraph then. Figure 4.6 shows how word proportions are calculated to output multiple topics for a given document.

Approach:

Like earlier approaches, we concatenated review summary and review text for each review. Then, we concatenated all the reviews together. Then we cleaned this text to lower case, removed stop-words, tokenized, created bigrams and trigrams, and lemmatized. We used the python library gensim and its implementation of the LDA model and noted the topic for all reviews for an ASIN. We now matched the product title with the new review topic, thus obtained by checking for the number of common words between them.

Observations

We got about 16% precision (Refer section 2.4). This method gave good results and is faster than BIR as we compare the product description with a shorter 10-word review topic.

There are some aspects Topic Modeling does not cover like, there is no information about synonyms. For example, the reviews for a product earphones discuss its sound, and the word sound is present in feedback, but this approach doesn't detect it and puts it as hijacked reviews.

There are products with only descriptive reviews like "great", "good product", "five stars", and "does the job" which also do not give correct results.

Thus, even this method faced the same issues as BIR and gave similar accuracy. But this method was faster and provided more results than all earlier methods, as discussed in chapter 6.

5. SUPERVISED INVESTIGATION

In the previous chapter, we used unsupervised methods and focused on uncovering review hijacking in data. However, we got only $< 25\%$ precision, i.e., output had only $< 25\%$ product listings as hijacked. Again, these would return complete listings and did not detect each hijacked review. Hence, these methods did not provide desirable results.

Another issue with these methods is that they are word-based, and there is no information about the semantics of the words. Adopting word embeddings like Word2Vec [45] and GloVe [46] can potentially alleviate this problem. Another issue we faced was the manual inspection of the results. This form of inspection is very time-consuming and repetitive. Hence, we consider synthetic data labeling for our data and apply supervised learning methods.

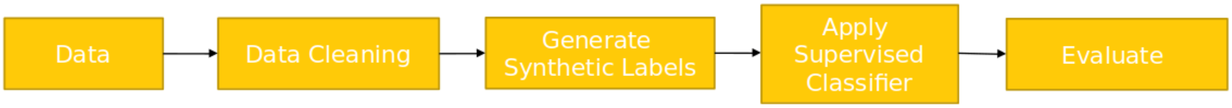


Figure 5.1: Plan for Supervised Investigation

Figure 5.1 shows our plan for the supervised investigation. We first perform data cleaning on dataset-2 (Refer section 3.2). Then we generate synthetic training data as we need labeled data for these supervised methods. We employ supervised classifier models and then evaluate our method using Accuracy, Precision-Recall curve and ROC curve. (These metrics are discussed in section 2.4).

5.1 Data Cleaning

We used Amazon dataset-2 (Refer section 3.2) for this task. For Data cleaning, we dropped unnecessary columns like 'also_buy', 'also_view', 'similar_item', 'price', 'date', 'details', 'tech1', 'tech2', 'fit', etc., and used only the columns shared here. We removed rows with no review text.

We removed the products with fewer than five reviews as such products are not very useful for our methods. We, later on, moved to 5-core reviews for the same reason. We also dropped some unformatted rows.

5.2 Synthetic Label Generation

Labels are required to train supervised learning algorithms. We would label our data as "RELATED" (= 0) and "UNRELATED" (= 1). Ideally, these labels should be generated manually by SMEs (Subject Matter Experts) or crowd-sourcing. But this form of manual work is expensive and time-consuming. So, we moved to other algorithmic methods to obtain not-so-accurate synthetic labels for our data.

5.2.1 Weak Supervision

Weak supervision takes an approach where we leverage the concepts of data programming [47] to unify and model multiple sources of weak labels to create a strong label. We obtain weak labels via heuristic rules, distant supervision techniques, keyword and/or pattern matches, third-party models, noisy labels from crowd workers, weak classifiers and more.

We started with using Snorkel [48] for this task. Snorkel is a system that combines various sources of weak supervision. Then, it learns a generative model to apply labels to a dataset programmatically. In the paper on Snorkel, Ratner et al. introduce the concept of Labeling Functions (LFs) as black box snippets of code written by SMEs that are in turn used to label subsets of unlabeled data. Each labeling function is a weak label generator based on the methods mentioned above for obtaining weak labels. A single LF can produce less-than-ideal training labels. Hence, Snorkel learns a generative model to combine outputs of multiple LFs and generates probabilistic labels.

Approach

We cleaned the data like in earlier methods. Then we performed the following labeling functions to label the data.

We formed two labeling functions for forming the weak labels :

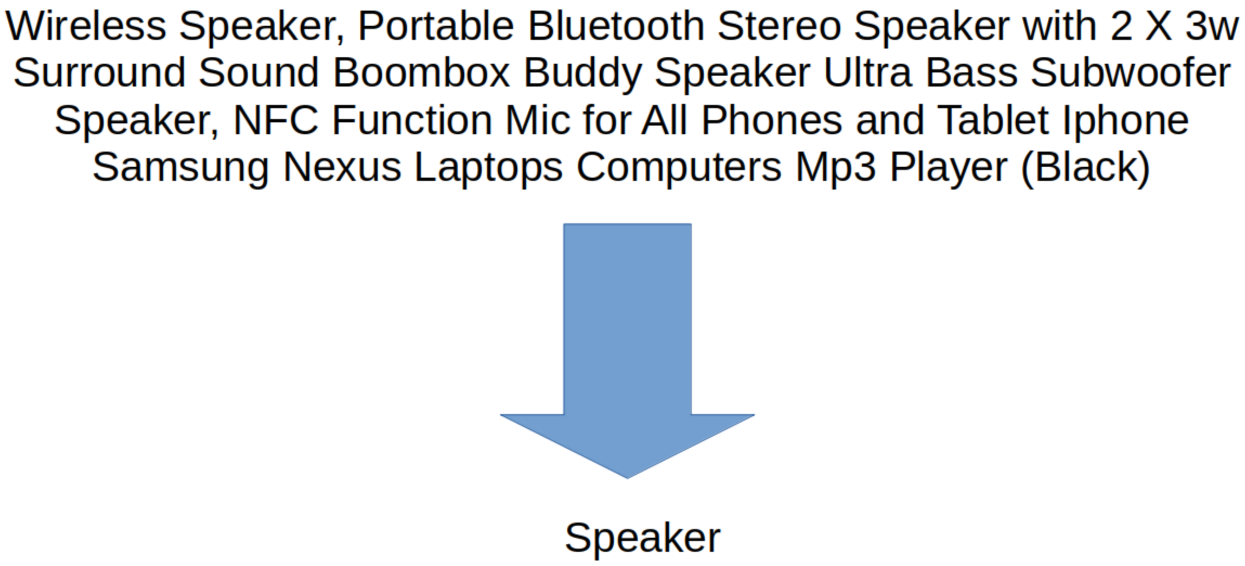

Wireless Speaker, Portable Bluetooth Stereo Speaker with 2 X 3w Surround Sound Boombox Buddy Speaker Ultra Bass Subwoofer Speaker, NFC Function Mic for All Phones and Tablet Iphone Samsung Nexus Laptops Computers Mp3 Player (Black)
Speaker

Figure 5.2: Actual Product Name Extraction from Product Title

- Labeling Function to determine the actual product name from the product title on Amazon. As we found the product titles very long and descriptive, we would identify the product names from the product title and search for them in the reviews. If the product name such obtained is in the reviews, then the review is labeled RELATED, else it is UNRELATED.

  We observed that the product name is generally one or two words. It is two words before words like ‘for’, ‘with’, ‘by’, ‘compatible with’, etc. or before symbols like ‘-’, ‘|’, ‘–’, etc. As we see from figure 5.2, the product name is just before the word ‘with’. This method provided the product names very efficiently for a lot of products.

- Labeling function for query expansion. We tried using synonyms for the product name from the NLTK wordnet, and similar to the above LF, the reviews were labeled RELATED and UNRELATED. This function did not perform very well for some of the data. For example, if a product is earphones, wordnet would not have proper synonyms for these words, and even the reviewers not use accurate synonyms like ’headphones’ for ‘earphones’

Observations

There were two issues that we faced in this method. First, Snorkel is buggy and so could not help labeling extensively. The process would die out or require restarting the kernel frequently. Second, Not all reviews have just product names or their synonyms in the title. For example, for a music player, “Sound was ok” is a valid review, while our methods detect it as invalid. This inaccuracy led to fewer results and was not very efficient labeling.

Labeling Functions are important for weak supervision. We did not have proper labeling functions to label the reviews effectively as none of the labeling functions could capture the context well.

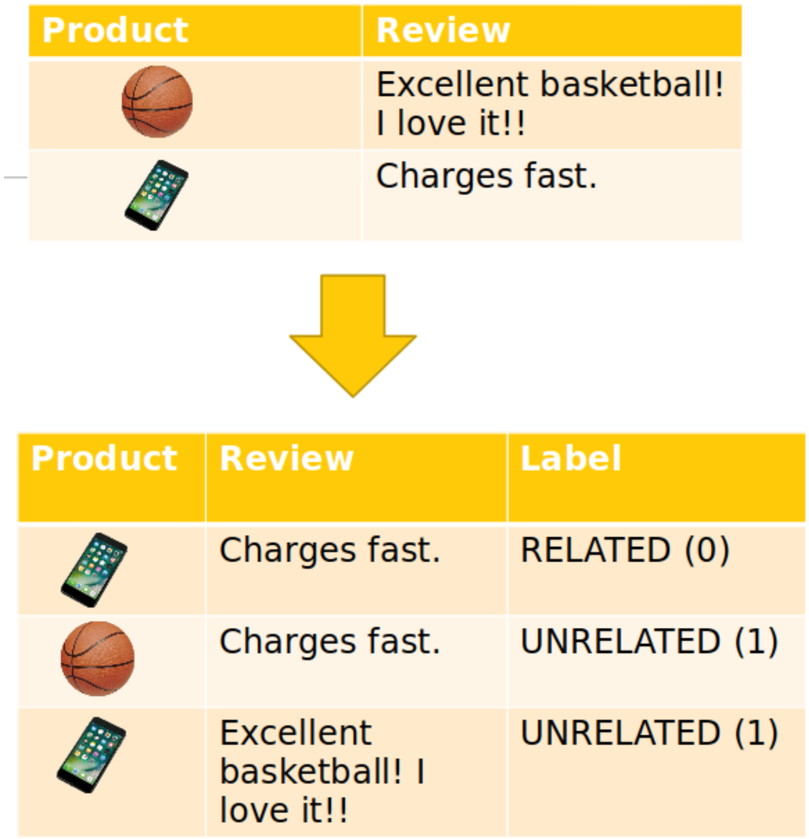
5.2.2 Swapping Reviews Among Products

We decided to swap reviews among a pair of distinct products to obtain unrelated data. For example, as seen from figure 5.3, we have a basketball and phone and their associated reviews. We swap reviews among the products to generate UNRELATED (1) data. We take all reviews associated with the data are RELATED (0). We know that our data has some hijacked reviews, so these related data may not always be related, but this percentage is tiny ($< 0.01\%$). Hence, we will tolerate this noise for synthetic label creation.

But, randomly switching product titles and review texts will not work. There are similar products in the dataset, and reviews of those similar products might not be unrelated. For example, if we have Samsung mobile covers from two different brands as products, their feedbacks will still be related or on-topic. Hence, we propose two methods for finding pairs of dissimilar products for review swapping:

1. Inter-category product swapping

2. Swapping based on Jaccard Similarity (Intra-Category Product Swapping)



Product	Review
🏀	Excellent basketball! I love it!!
📱	Charges fast.

Product	Review	Label
📱	Charges fast.	RELATED (0)
🏀	Charges fast.	UNRELATED (1)
📱	Excellent basketball! I love it!!	UNRELATED (1)

Figure 5.3: Generation of Synthetic Label and Data by Swapping Reviews among Dissimilar Products

1. Inter-Category Product Swapping

We have products in the dataset separated into different categories like Beauty, Clothing, Electronics, Cellphones & Accessories, etc. We take products from one category and the reviews from another for UNRELATED reviews. We take the reviews belonging to the same product to be RELATED.

Approach

We use basic pandas dataframe methods for this swapping method. We load the data of two categories, taking the product title and description of one and the review summary and text of the other. Then, we randomly assigned reviews to products to obtain UNRELATED review data.

Now, like before, we assumed that the products do not have any hijacked reviews and used product review pairs in the original file as RELATED data. We then shuffled them up and split them into training, validation and test set. We would call this dataset *Inter-Category product data*.

2. Swapping based on Jaccard Similarity (Intra-Category Product Swapping)

The earlier method only produces data when products of different categories have hijacked reviews. For hijacking occurring within a product category, we use Jaccard distance.

Jaccard index or Jaccard similarity coefficient is a statistic that measures similarities between two sample sets. Its mathematical definition is as below.

$$J(A_1, A_2) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} = \frac{|A_1 \cap A_2|}{|A_1| + |A_2| - |A_1 \cap A_2|}$$

We use this method for products in the same category to generate intra-category product data.

Approach

We converted product titles for each product into TF-IDF feature matrices and found pairwise Jaccard distances between them. Then, we formed product pairs (A_1, A_2) with Jaccard distance 0. Then, we took the product title and description of one product A_1, and the review summary and text of another product A_2 and labeled this as UNRELATED. Similarly, we took the product title and description of A_2 and a review of A_1 as UNRELATED. For RELATED labels, we took the product title & description, and the review summary & text of A_1, and likewise for A_2 to get another set of RELATED data.

As we took Jaccard value 0 as our threshold, the UNRELATED reviews are decently unrelated. For RELATED data, we assumed that reviews are related to its product. This relatedness may not be true for hijacked reviews in the data, but we accept this noise.

5.2.3 Comparison between Methods for Synthetic Data Generation

Weak supervision did not provide acceptable labels, and hence we discarded this approach. For inter-product category data, it simply depends on product categories. We calculate Jaccard distance using TF-IDF vectors. So it depends on the words in the product title. This dependence associates the products that use different words in the product title. We can use Jaccard distance for inter-category and intra-category data. Inter-category data can develop expertise on one product category and can detect hijacked reviews efficiently for that particular category.

The data generated from these synthetic methods are noisy. The RELATED labeling is improper for hijacked reviews or reviews added mistakenly for wrong products. Again, if some feedbacks are generic like “Good product”, these reviews will be labeled RELATED or UNRELATED, depending on the product pair and not on the actual relation. Even though the data generated is quite noisy, this noise is sufficiently low for our models to work.

5.3 Supervised Learning Algorithms

Supervised learning is the most common form of a machine learning (ML) task. Here, the ML model trains on the labeled data, where each input has a specific output. For example, given a set of images, some containing trees and others not, it learns which images are trees and which are not. In a supervised learning setup, we divide the data into train set, validation set, and test set. The supervised model trains on the train set, confirm validity on the validation set, and provides results on the test set.

We use supervised deep learning approaches to fix two problems. First, unsupervised learning approaches provided low precision. Hence, we would now use better and modern algorithms to get better results. Second, they did not use the semantic information of the words. Hence, we use GloVe word embeddings for Siamese LSTM and the BERT sequence pair classifier uses BERT representations by default.

We use the following two approaches:

1. Pair-wise classification using Siamese LSTM Networks

2. Finetuning BERT on Sequence pair Classification Task

5.3.1 LSTM Siamese Network

We have always been using LSTM networks for text-based machine learning tasks. A Siamese network, also called a twin neural network, uses the same weights parallel in tandem on two inputs to return output based on relation or distance between them [49]. Generally, they are used to check the similarity between images and texts. We use LSTM Siamese networks to compare sentence pairs and determine if they are similar or not. We would use the same approach for our data as well.

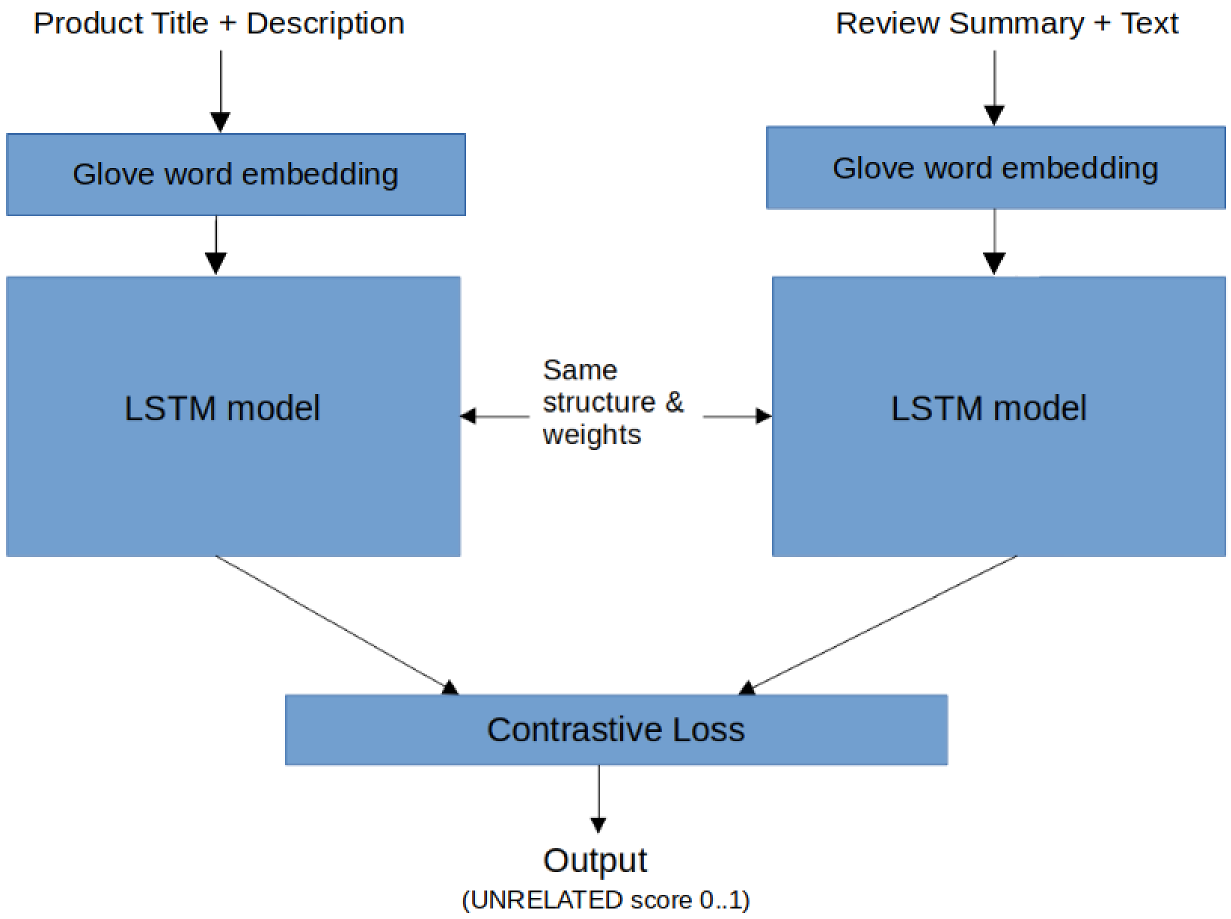


Figure 5.4: Siamese LSTM Network for Review hijacking Detection

Approach

We concatenated the product title and product description for our first input. For our second input, we concatenated the review summary and review text. Then, we tokenized our inputs and converted them into sequences. Then we used 300-dimensional GloVe [46] embeddings and formed an embedding metric for our tokens. We get two embedding matrices for both our inputs, and we feed them into our LSTM network.

We implemented a siamese network, as shown in figure 5.4. This network configuration is also called MaLSTM (Manhatten LSTM). We use two LSTM networks with 64 nodes and two layers each. We take the two input embeddings and calculate binary cross-entropy loss with Adam optimizer to get the result.

5.3.2 BERT Sequence Pair Classifier

We know that BERT has provided state-of-the-art (SOTA) results for various NLP tasks such as Question-Answering (QA), Natural Language Inference (NLI), etc. Hence, we attempt to model it for our project as well.

Google developed Bidirectional Encoder Representations from Transformers (BERT) in 2018, and it has been used widely in all NLP tasks since then. It is a Transformer-based machine learning tool for pre-training unlabeled text for Natural Language Processing (NLP) tasks.

BERT provides a deep bidirectional representation, conditioned on text in both directions, left-to-right and right-to-left. It understands the full context of a word based on the ones before and after it. LSTM used GloVe word-embeddings that provide semantic of the words, but not context like in BERT. Thus, we expect this method to perform better than the previous Siamese LSTM method.

Approach

Our model is prepared from the BERT BASE model (bert_12_768_12) from GluonNLP [51].

First, we load the BERT model using the model API. This model is pre-trained using a corpus of books and the English Wikipedia. Then, we add a layer on top for classification, as shown in

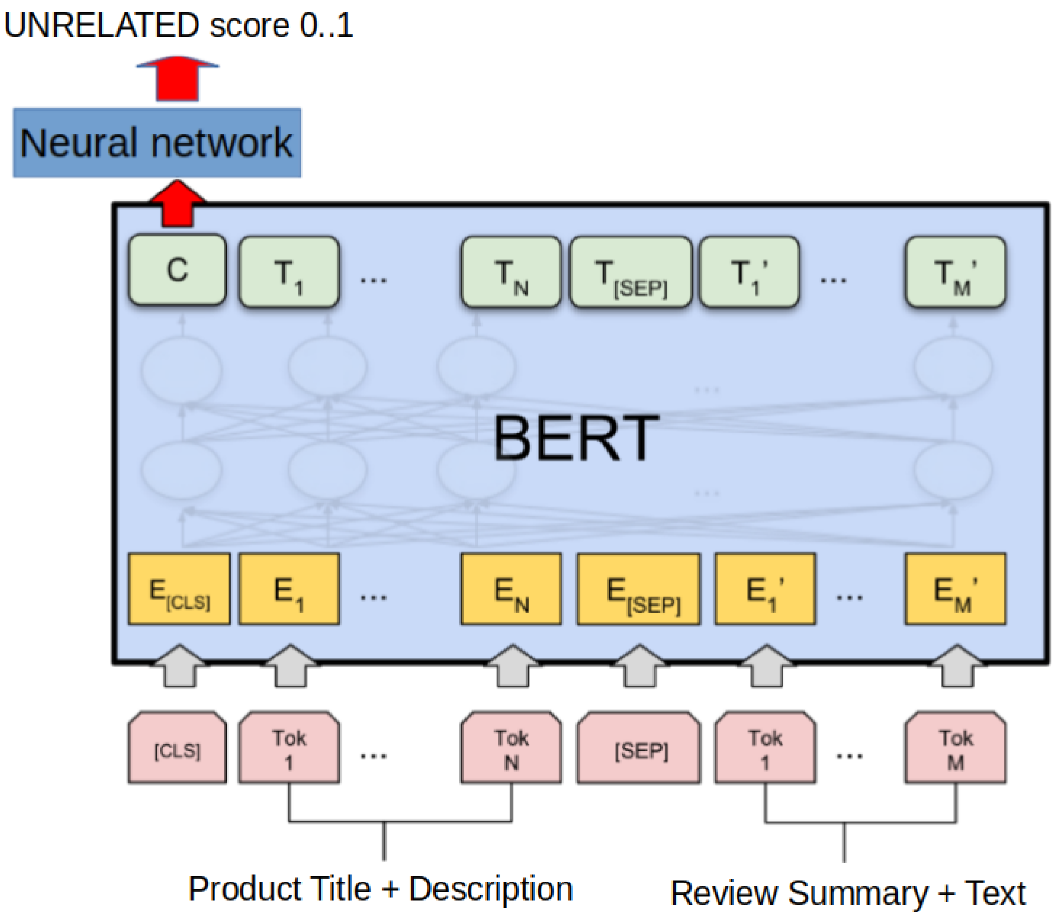


Figure 5.5: BERT Model for Review Hijacking Detection

figure 5.5. We use adam optimizer for optimizing this classification layer.

Now we form the sentence pairs for classification. Like the previous method, the first sentence is a concatenation of product title and product description. We obtain the second sentence by concatenating the review summary/headline and the review text. We then tokenize the sentences, insert [CLS] at the start, insert [SEP] at the end and between both the sentences, and generate segment ids to specify if a token belongs to the first sentence or the second one.

We now run the BERT fine-tuning with these sequences as inputs. We get the output as an *UNRELATED score* between 0 and 1.

6. RESULTS AND CONCLUSIONS

6.1 Results for Unsupervised Models

Model	Precision	Number of results
BIR	17%	3069
TF-IDF / BM25	23%	1134
BIR with NER	12%	630
Topic modeling	16%	6993

Table 6.1: Results for Unsupervised Learning Algorithms

Table 6.1 shows the results we have obtained from unsupervised learning algorithms. Here, we can see that results have a very low accuracy of < 25%. The models can detect some patterns but insufficient to predict unrelated reviews reliably.

There might be various reasons for such results. One of the reasons might be no semantic information. There were no word-embeddings or sentence embeddings applied, and hence the methods relied on only word counts. This absence of semantics resulted in these techniques being more of ablation tasks. They provide a glimpse of some hijacked reviews present in our data but do not provide enough results.

We also see that BIR (Boolean method of information retrieval) and Topic modeling provided more results (number of hijacked review listings) than the other methods used. TF-IDF has higher accuracy but fewer samples because TF-IDF considered frequent words like "iPhone" and "case" as unimportant even though these words were crucial to detection. BIR and Topic modeling performed similarly for accuracy. BIR (Boolean method of information retrieval) with NER (Named Entity Recognition) provided the worst results of all methods due to bugs in TextBlob and other NER methods.

Topic modeling got the maximum number of results from the data, and hence we would consider this as the best performing method of the lot. It is even faster to execute than other methods.

6.2 Results for Supervised Models

Model	Syn. Data	Accuracy	ROC result
Siamese LSTM Network	Jaccard dist.	0.823	0.770
	Inter-category	0.885	0.910
BERT Sequence pair classifier	Jaccard dist.	0.916	0.948
	Inter-category	0.965	0.993

Table 6.2: Results for Supervised Learning Algorithms

Table 6.2 shows us results obtained from supervised learning Siamese LSTM and BERT for our data. The results are better than the unsupervised methods as we have used semantic information of these texts. Again, LSTM and BERT are better and modern classification models than the earlier unsupervised methods.

We should note that the results are on synthetic data, and the data is noisily labeled. So the results might not be as accurate as they appear here and should be taken with a grain of salt.

Following are some of the precision-recall curves and ROC curves obtained from the supervised learning methods :

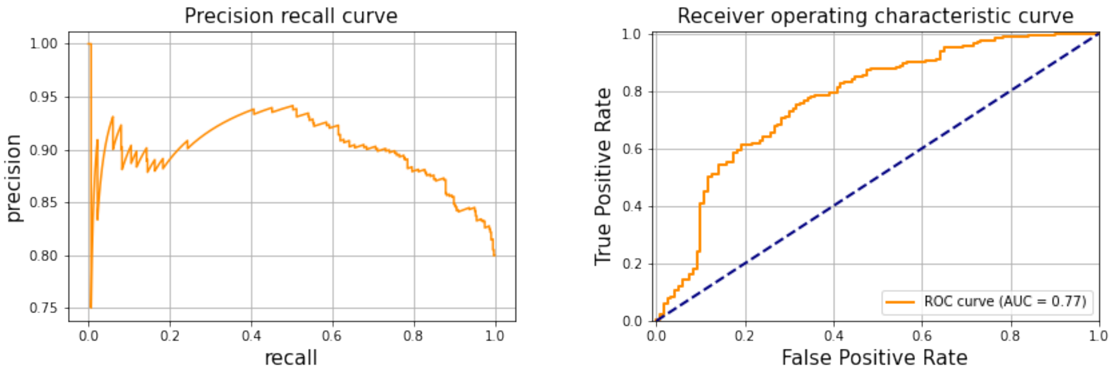


Figure 6.1: ROC and PR curve for Siamese LSTM network run on Intra-category data (Jaccard distance)

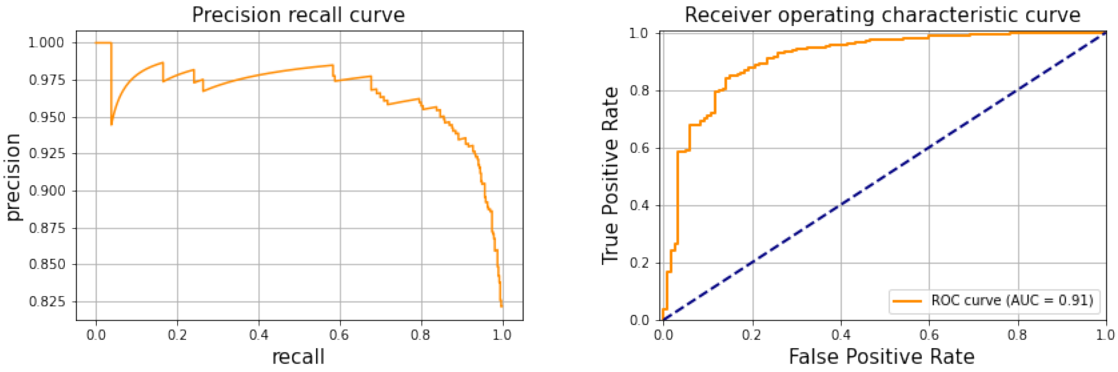


Figure 6.2: ROC and PR curve for Siamese LSTM network run on Inter-category data

We see from these figures 6.1, 6.2, 6.3, and 6.4 that BERT fine-tuning performs better than Siamese LSTM in both datasets. Siamese LSTM used GloVe embedding, and BERT used its bi-direction embedding. LSTM could use semantic knowledge of words, while BERT used right and

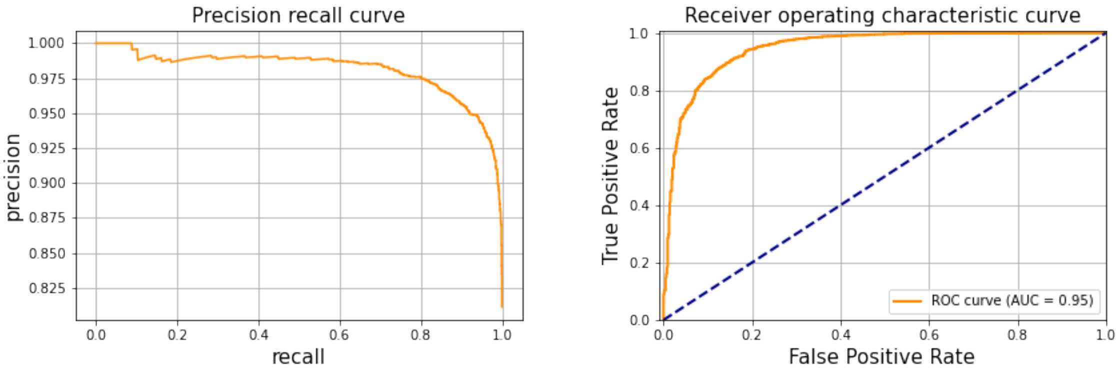


Figure 6.3: ROC and PR curve for BERT seq. pair classifier run on Intra-category data (Jaccard distance)

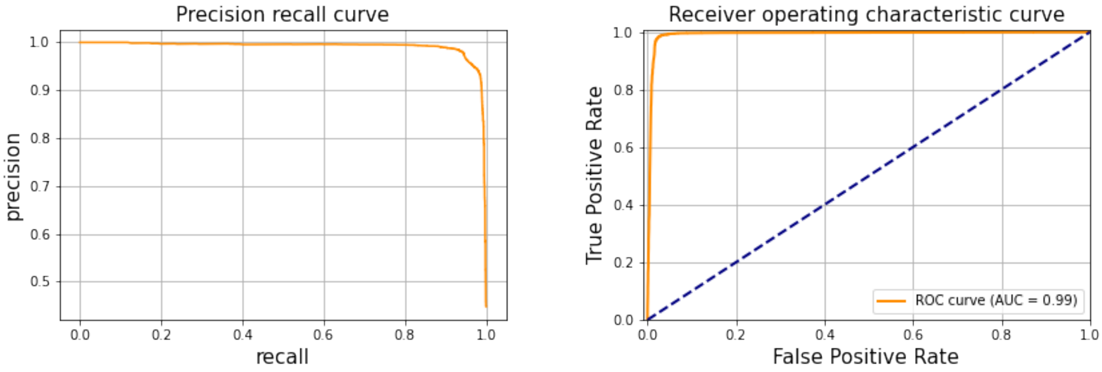


Figure 6.4: ROC and PR curve for BERT seq. pair classifier run on Inter-category data

left semantics of phrases and even sentences. Thus, precision and recall are higher for the BERT Fine-tuning model.

We also see that both methods perform better on the inter-category dataset than the intra-category (calculated using Jaccard distance) dataset. We obtain UNRELATED reviews in the inter-category dataset by taking products from one category and review texts from another. Hence, models trained on this dataset can learn product features of one category at a time and develop expertise in that category. Thus it performs better than Jaccard distance data, where the items can be from one or more categories. This labeling method does not work if we have products from only one class.

6.3 Evaluation on Ground Truth Hijacked Listings

We ran our BERT sequence pair classifier algorithm on inter-category data on a sample dataset of 31K products (6.5 M reviews) with the original product review pair intact. These 31K products were held out and not used during the training.

Then, we calculated the average review score for each product as follows:

$$\begin{aligned} Avg.\ Review\ Score &= \frac{UNRELATED\ score\ review_1 + ... + UNRELATED\ score\ review_N}{Total\ number\ of\ reviews\ of\ a\ product(N)} \\ &= \frac{Sum\ of\ UNRELATED\ scores\ of\ all\ reviews\ of\ a\ product}{Total\ number\ of\ reviews\ of\ the\ product} \end{aligned}$$

We manually checked a small sample of 200+ products with an average review score > 0.5 and found 99.95% of the listings containing unrelated or hijacked reviews. This high percentage shows that our model can predict hijacked listing with excellent accuracy.

From the graphs in figure 6.5, we can see that there are very few products with a higher average review score. About 99% products have review scores below 0.3, reinforcing our initial assumption about skewed class distribution.

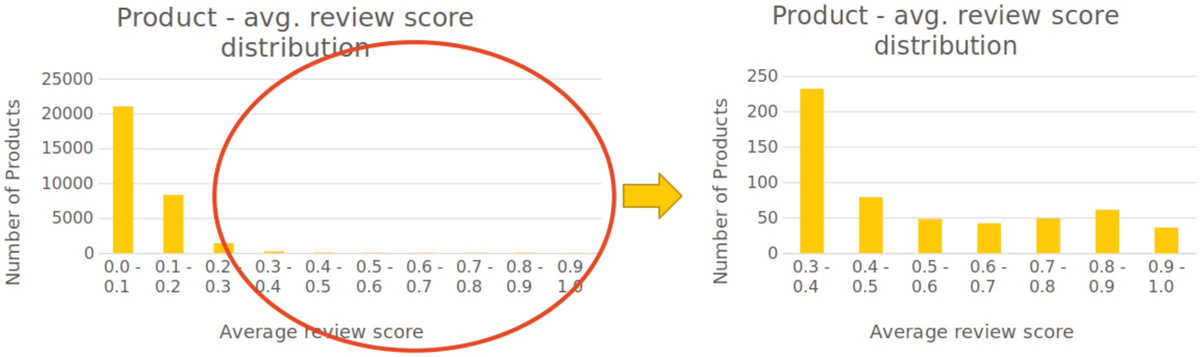


Figure 6.5: Average Review Score vs. Number of Products

We shall see 3 sample products and their distribution of the UNRELATED score for each review.

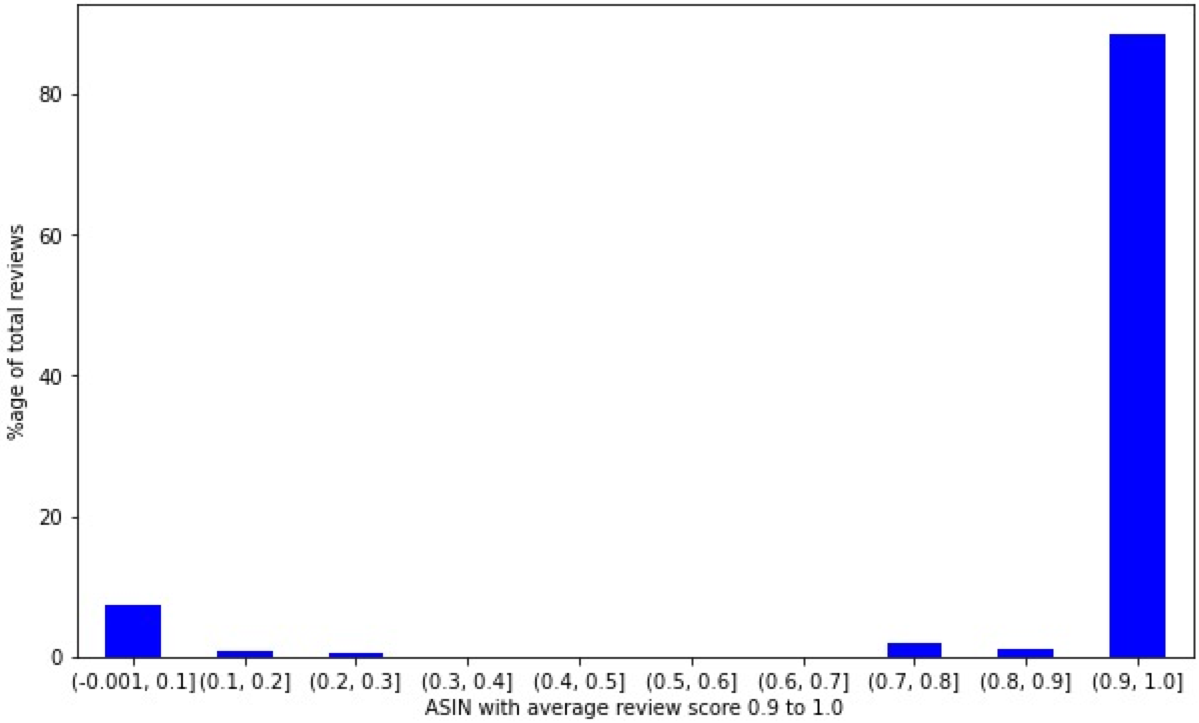


Figure 6.6: UNRELATED Review Score Distribution for Product-1

Figure 6.6 is the UNRELATED review score distribution of each review for product-1. Product-1 has an average review score of 0.9 to 1.0. We can see from the distribution that most reviews have a high UNRELATED score (> 0.9).

We manually inspect these reviews with a high UNRELATED score (> 0.9). We observe that these reviews are unrelated, and hence this product is a sample of review hijacking.

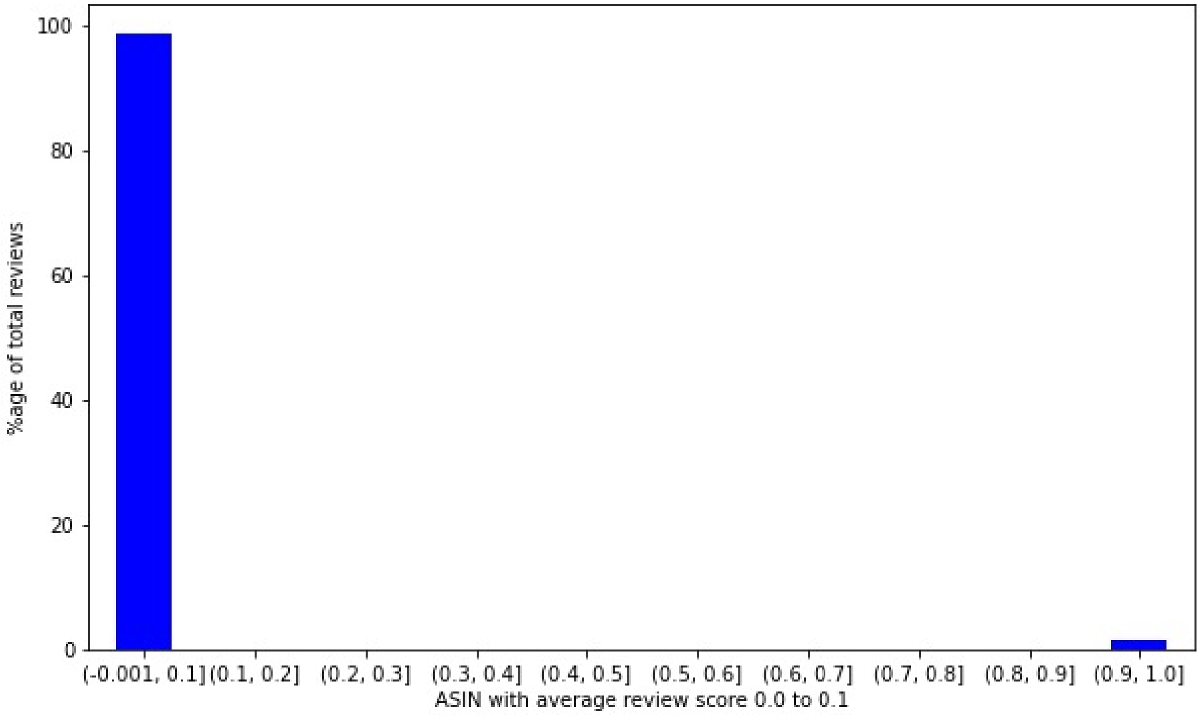


Figure 6.7: UNRELATED Review Score Distribution for Product-2

Figure 6.7 is the UNRELATED review score distribution of each review for product-2. Product-2 has an average review score of 0.0 to 0.1. We can see from the distribution that most reviews have a low UNRELATED score (< 0.1), and a few have a high score (> 0.9).

We manually inspect these reviews with a high UNRELATED score (> 0.9). We observe that these reviews are either misclassified as unrelated or do not have enough information to determine the label. These reviews are sometimes generic, like “Great Product!”. Thus, this product is not an example of review hijacking.

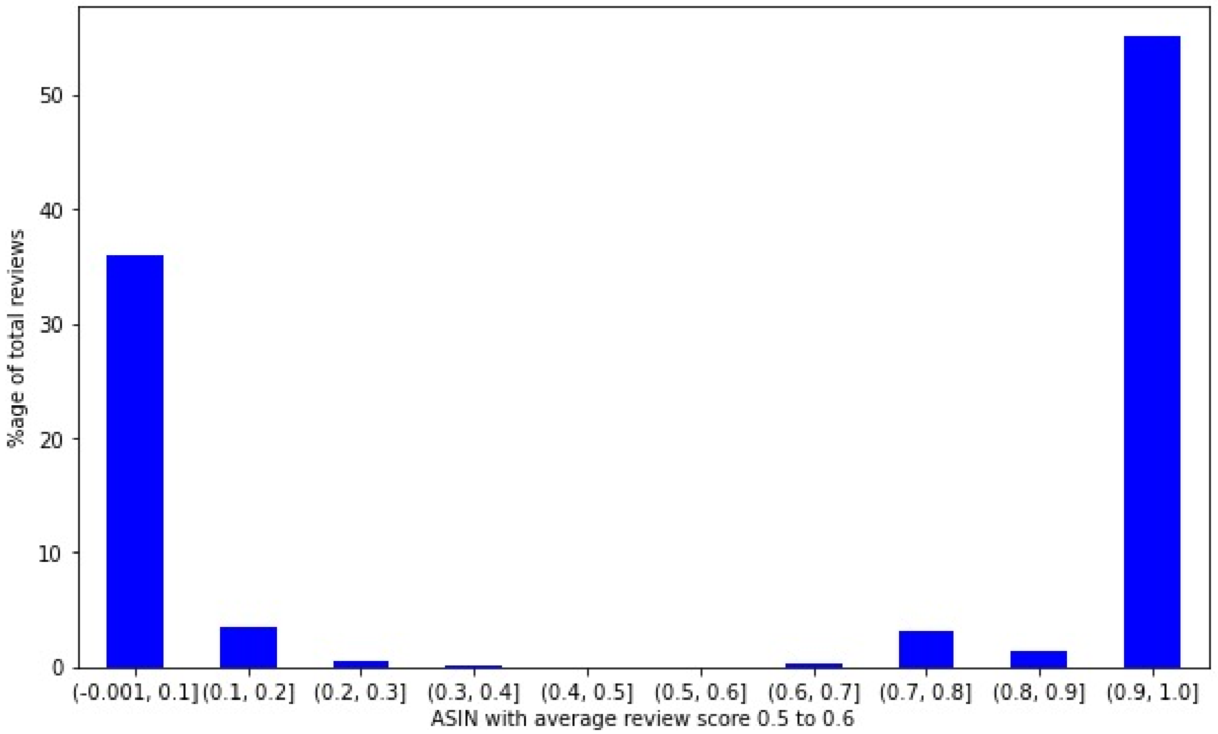


Figure 6.8: UNRELATED Review Score Distribution for Product-3

Figure 6.8 is the UNRELATED review score distribution of each review for product-3. Product-3 has an average review score of 0.5 to 0.6. We can see from the distribution that about 55% of the reviews have a high UNRELATED score, while 35% reviews have a low UNRELATED score.

We manually inspect and observe these reviews with a high UNRELATED score (> 0.9) and observe that these reviews are unrelated. We also inspect the reviews with low UNRELATED scores (< 0.1 and < 0.2) and observe that most reviews are related to the product. As some of the reviews are unrelated, we determine this product is also a sample of review hijacking.

Another important observation from all these plots is that our model still predicts most reviews with high confidence (> 0.9 or < 0.1), even if the average review score is mid-range ≈ 0.55 for product-3. We observe that we can also label ASINs with some of the reviews as unrelated, which solves the challenge of seller techniques, as discussed in section 1.4.

6.4 Conclusion and Summary

We shed light on a new review manipulation technique called Review Hijacking. We discussed techniques sellers used for this fraud and shared some news reports for review hijacking. We observed that even though it is well-reported, there has been little to no academic research, and hence we aimed to kick start this domain. We utilized datasets consisting of Amazon product reviews for this.

As there is no labeled data, we first started with an unsupervised investigation of Review Hijacking. We tweaked unsupervised Information Retrieval (IR) algorithms to obtain insights and the extent of review hijacking. We observed having a skewed dataset with few samples ($< 0.01\%$) of hijacked reviews.

We then planned to explore supervised learning methods. To obtain ground truth for these approaches and resolve the skewed data issue, we proposed novel methods for synthetic label and data generation by swapping the reviews of a product with reviews on an unrelated product. We then applied supervised classifiers like the Siamese LSTM network and BERT sentence pair classifier to obtain results.

Both the Siamese LSTM network and BERT sentence pair classifier provided excellent results on synthetic data, but the BERT sentence pair classifier provided better results. We evaluated this method on real-world data and observed that it provided excellent results on that too. Hence, we propose the BERT sentence pair classifier as an efficient detection algorithm for this task.

7. IDEAS FOR FUTURE WORK

We can further expand our work to get better results.

7.1 Data

We can find better label generation methods as synthetic label generation provides in-accurate noisy data. We can use crowd-sourced data, like Amazon Mechanical Turk (MTurk), labeling for more accurate labeling. This method requires some funding. Well-labeled data might provide better results.

We can perform web-scraping for recent data. We have data until 2018, and the first news report for review hijacking was also in 2018. If we get data for 2019 and 2020, we might get more review hijacking samples. Apart from that, we may also obtain more strategies applied by the sellers for review hijacking. Newer data will also help us show the impact on recent products.

7.2 Methods / Algorithms

We can expand our work to use the images attached to the product and shared by the customers. Thus, we can include Computer Vision tasks as well.

We can perform newer state-of-the-art transformer architecture, like T5, DeBERTa, RoBERTa, etc., on this task and check performance.

In this thesis, we have only fine-tuned BERT. The model is pre-trained on the concatenated corpus of BooksCorpus and English Wikipedia with all lowercase letters. We can further progress on the BERT model by pre-training on Amazon catalog data, changing vocabulary to specific data.

REFERENCES

[1] J. Schlosser, “9 out of 10 consumers read reviews.” Content Customs, January 2021.

[2] N. Nguyen, “Here’s another kind of review fraud happening on amazon.” BuzzFeed News, May 2018.

[3] J. Swearingen, “Hijacked reviews on amazon can trick shoppers.” Consumer Reports, August 2019.

[4] “Variation relationships overview.” Seller Central Amazon.

[5] K. Merton, “The world’s top online marketplaces 2020.” Web Retailer, 2020.

[6] R. Murphy, “Local consumer review survey 2020.” Bright Local, December 2020.

[7] D. Kaemingk, “Online reviews statistics to know in 2021.” Qualtrics, October 2020.

[8] “2018 reviewtrackers online reviews stats and survey.” Review Trackers, 2019.

[9] R. P. II, “Should you buy google reviews? read this.” Review Trackers, January 2021.

[10] S. Ovide, “How fake reviews hurt us and amazon.” NY Times, November 2020.

[11] H. Walsh, “How ebay’s review system is promoting fake, counterfeit and even dangerous products.” Which, March 2020.

[12] L. James, “Fake yelp reviews: Defending your business from online falsehood.” Reputation Sciences, May 2020.

[13] J. Wattles, “Amazon sues more than 1,000 sellers of ’fake’ product reviews.” CNN Money, October 2015.

[14] I. Lee, “Can you trust that amazon review? 42monitor says.” Chicago Tribune, September 2020.

[15] “Ftc brings first case challenging fake paid reviews on an independent retail website.” FTC.gov, February 2019.

[16] “Amazon fake reviews: a problem and a 5-star industry.” ReviewMeta.

[17] E. Dwoskin and C. Timberg, “How merchants use facebook to flood amazon with fake reviews.” Washington Post, April 2018.

[18] “I write fake reviews for yelp and other service sites.” Rediff, April 2016.

[19] K. Schoolov, “Amazon is filled with fake reviews and it’s getting harder to spot them.” CNBC, September 2020.

[20] Z. Bernard, “Some amazon employees are reportedly accepting cash bribes from online sellers to delete negative product reviews.” Business Insider, September 2018.

[21] I. Steiner, “Click fraud could prove costly to marketplace sellers.” E-commerce bytes, May 2019.

[22] D. Fuchs, “Amazon’s quest for more, cheaper products has resulted in a flea market of fakes.” Y Combinator, November 2019.

[23] “Ship of theseus.” Wikipedia.

[24] G. Sterling, “Review fraud: Hijacked amazon reviews a big problem says consumer reports.” Search Engine Land, August 2019.

[25] “Swapped product listings on amazon - web applications stack exchange.” Webapps Stackexchange, 2019.

[26] T. B. Lee, “Amazon still hasn’t fixed its problem with bait-and-switch reviews.” Arstechnica, December 2020.

[27] B. Lynch, “’bait and switch’ amazon reviews cause skeptical consumers.” Customer Contact Week Digital, January 2021.

[28] “Let’s stop hijacked reviews!.” Consumer Reports, 2019.

[29] “#stopreviewhijacking on twitter.” Twitter, 2019.

[30] J. Dzieza, “Even amazon’s own products are getting hijacked by imposter sellers.” The Verge, August 2019.

[31] A. Mukherjee, B. Liu, and N. Glance, “Spotting fake reviewer groups in consumer reviews,” in *Proceedings of the 21st international conference on World Wide Web*, pp. 191–200, 2012.

[32] A. Mukherjee, V. Venkataraman, B. Liu, N. Glance, *et al.*, “Fake review detection: Classification and analysis of real and pseudo reviews,” *UIC-CS-03-2013. Technical Report*, 2013.

[33] P. Kaghazgaran, J. Caverlee, and M. Alfifi, “Behavioral analysis of review fraud: Linking malicious crowdsourcing to amazon and beyond,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, 2017.

[34] D. Higgins, J. Burstein, and Y. Attali, “Identifying off-topic student essays without topic-specific training data,” *Natural Language Engineering*, vol. 12, no. 2, p. 145, 2006.

[35] A. Louis and D. Higgins, “Off-topic essay detection using short prompt texts,” 2010.

[36] M. Rei and R. Cummins, “Sentence similarity measures for fine-grained estimation of topical relevance in learner essays,” *arXiv preprint arXiv:1606.03144*, 2016.

[37] S. Ryu, S. Kim, J. Choi, H. Yu, and G. G. Lee, “Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems,” *Pattern Recognition Letters*, vol. 88, pp. 26–32, 2017.

[38] A. Hanselowski, A. PVS, B. Schiller, F. Caspelherr, D. Chaudhuri, C. M. Meyer, and I. Gurevych, “A retrospective analysis of the fake news challenge stance detection task,” *arXiv preprint arXiv:1806.05180*, 2018.

[39] “Amazon customer reviews dataset.” S3 Amazon-AWS, 2015.

[40] “What are upcs, eans, isbns. and asins?.” Amazon.com.

[41] “Amazon standard identification number.” Wikipedia.

[42] J. Ni, J. Li, and J. McAuley, “Justifying recommendations using distantly-labeled reviews and fine-grained aspects,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

[43] E. Lancaster, F.W.; Fayen, “Information retrieval on-line,” 1973.

[44] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta, Malta), pp. 45–50, ELRA, May 2010.

[45] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013.

[46] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[47] A. Ratner, C. D. Sa, S. Wu, D. Selsam, and C. Ré, “Data programming: Creating large training sets, quickly,” 2017.

[48] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, “Snorkel,” *Proceedings of the VLDB Endowment*, vol. 11, p. 269–282, Nov 2017.

[49] D. Chicco, *Siamese Neural Networks: An Overview*, pp. 73–94. New York, NY: Springer US, 2021.

[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019.

[51] “Fine-tuning sentence pair classification with bert.” GluonNLP.

[52] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *CoRR*, vol. abs/1908.10084, 2019.