Review Hijacking in Online Shops

Stefan Octavian Simionescu
University of Twente P.O. Box 217, 7500AE Enschede The Netherlands
s.simionescu@student.utwente.nl

Abstract

Customer reviews and ratings are critical decision-making tools for online customers in the rapidly growing e-commerce space. However, review hijacking, opinion spamming, and suppressing negative reviews are a few of many deceitful tactics that can undermine the legitimacy of the review ecosystem. Review hijacking is the practice of replacing a popular product with a different one to increase sales and popularity. The system does not operate as if it were a novel listing, as it is merely a modification of an existing product. By doing this, the product advertised ends up using unrelated reviews, and because of the number of reviews, it usually ends up showing more often in its category. This paper addresses the issue of review hijacking or review-reuse by utilizing machine learning models to determine whether a review aligns with the product. During the research, a RoBERTa model and a base BERT model for text classification were used with the aim of achieving higher accuracy in review hijacking detection with the RoBERTa model. Moreover, this paper presents a visualization concept, utilizing the implemented machine learning model, that exhibits superior performance.

Keywords— hijacking reviews, online shopping, machine learning

1 Introduction

Reviews and ratings are particularly important in facilitating purchasing decisions by online shoppers [1, 2, 3, 4]; customers are more likely to rely on others' experiences than on their own thoughts on the seller's direct marketing [5]. Furthermore, the increased reliance on user reviews stems from their tendency to convey a subjective perspective. This elucidates whether and how a product aligns with a specific individual's preferences and usage context, while the information offered by the seller typically focuses on product attribute information [6]. Product choice, as well as the perceived trustworthiness of an online store, is influenced by consumer reviews [7]. 79% of individuals read reviews in 2020 [8], while in 2013, 80% of individuals read reviews [9]. Therefore, online reviews remain a relevant tool in facilitating customers in their purchasing decision. Moreover, firms are motivated to watch and manipulate online product reviews on their websites or third-party platforms to alter both the quantity and the quality of the reviews to influence consumer perceptions [10, 11]. There are many review manipulation techniques, such as fake reviews, paid reviews, boosted reviews, hijacked reviews [12].

Review hijacking is a recently appearing unethical practice of substituting the details of a popular, positive-reviewed product with those of a completely unrelated one [13, 14, 15]. In this way, the newly advertised product uses the reputation of the prior product to increase its sales in an untrustworthy manner [16]. This deceitful operation tricks both customers and the ranking system, favouring goods with numerous reviews and high ratings. Review hijacking gives the seller an unfair advantage [17].

Regarding the matter of fake reviews, there exists a substantial body of research [18, 19, 20, 21, 22, 23]. Contrariwise, review hijacking is not a well-studied manipulation tactic. Therefore, this research project addresses the issue of review hijacking. Hijacked reviews are genuine reviews, with verified purchase status, but from other products. Hence, the detection of fake reviews is ineffective for these reused reviews. To identify hijacked reviews, it will be necessary to employ semantic analysis to determine the correlation between the product title, description and reviews. One alternative approach to determining correlation between two pieces of text is to employ a Twin neural network [24, 25]. Although, BERT (Bidirectional Encoder Representations from Transformers) [26] yields higher accuracy when trained with synthetic data for detecting hijacked reviews [14]. RoBERTa, stands for Robustly Optimized BERT Pretraining, is an improved version of BERT [27]. Given that RoBERTa is a novel and enhanced version of BERT, there is a reason to believe that it has the potential to deliver higher accuracy in the detection of hijacked reviews.

During this research, significant observations were made that can aid in the detection of review hijacking. Several of these observations pertain to the duration of the reviews, the country of origin and the information provided by the seller. Both RoBERTa and

BERT performed well when it came to detecting hijacked reviews, with a slight advantage for RoBERTa. This will be further detailed throughout this paper. We conclude this paper by clarifying the implications of the findings and answering the research questions. By the end of the research, we will detail how we reached the conclusion that RoBERTa performed better than BERT for review hijacking and that a credibility score could be effectively communicated to a user interface through various visual components. To facilitate future research in this field, we offer valuable insights regarding the limitations of this study and its potential enhancement of review hijacking, as well as how future research can be based on these findings. Moreover, with this work, we contribute a dataset to the field of NLP in e-commerce to enable the design and evaluation of techniques for identifying review hijacking. This will enable improvements in fairness in e-commerce and ethical business practices.

2 Related Work

In the domain of review manipulation, there exists significant research, with the majority of it focusing on fake reviews [18, 19, 20, 21, 22, 23]. Fake reviews are false claims that are made by people or machines regarding the experience of a product [23, 15]. Different methods have been proposed to detect fake reviews, such as textual features [28, 29], supervised machine learning using BERT models [30], Bidirectional Long Short-Term Memory (Bi-LSTM) method [31], and many more [21]. There are also studies that demonstrate how well the RoBERTa language model performs in detecting fake reviews [32] [33][34]. Since text generation methods yield realistic reviews, it is difficult for humans to detect fake reviews nowadays. However, learning classifiers can detect fake reviews with almost perfect accuracy [23]. In this field of research, the methods implemented focus on textual analysis [29, 21], behavioural analysis [35, 36], reviewer credibility [15]. The evidence presented thus far supports the idea that these methods are used because fake reviews have specific characteristics that can be detected using various methods. Fake reviews are usually detected based on both the type of review and certain attributes that are not directly related to the content [21]. Therefore, there is a primary focus on detecting unusual patterns in detecting if a review reflects the genuine opinion of a customer or if it is fake.

On the contrary, researchers have not treated review hijacking in much detail. Regarding the subject of reused reviews, a Twin LSTM (Long Short-Term Memory) neural network and Bert-based classifier have been trained on synthetic data to identify hijacked reviews. Both approaches produced excellent results when compared to artificial data and numerous instances of hijacked reviews were discovered. The BERT-based classifier between 91% and 96% percent accuracy, compared to the Twin LSTM Network's 82% to 91%. The reviews were categorized into two distinct types, namely related and unrelated reviews. The method investigated included manual labour for annotating synthetic data for product-review pairs [14]. Research on this subject has been limited to simply identifying reused reviews, based on a model trained on synthetic data and calculating a suspiciousness score [14].

3 Research Objectives

This paper seeks to investigate how review hijacking can be identified and quantified. Moreover, it analyses how to improve on the performance of a BERT-based classifier for review hijacking using real reviews collected from Amazon as training data. The additional objective is to identify an effective means of conveying the legitimacy of a product, particularly in terms of review hijacking, to a potential user. The following research questions were posed and will be answered during this research:

(RQ_1): "To what extent does RoBERTa improve previously proposed BERT-based approaches for the task of review hijacking? [14]"

(RQ_2): "How can review hijacking be quantified?"

(RQ_3): "How can the likelihood of review hijacking be communicated effectively to users for products on e-commerce platforms?"

4 Methods

To carry out this research, a series of steps were undertaken, including the gathering of the dataset, the preprocessing and annotation of a dataset of product reviews from Amazon, as well as the training and evaluation. In this chapter, we will go into further detail about each of these steps.

4.1 Gathering the dataset

There is an already existing dataset composed of 32 hijacked products and 32 non-hijacked products that contains around 13k reviews for all 64 products. This existing dataset covers headphones products. The newly collected dataset contained 27 hijacked products and 26 non-hijacked products. The final dataset of all 117 products resulted in a total of 22,662 reviews. The products were meticulously selected manually from the Amazon website. During this phase, numerous observations were made. Almost all hijacked products contained unrelated reviews from books or rented DVDs. The majority of these unrelated reviews were in another language. These aspects will be discussed further in section 6. Apify's Amazon Review Scraper was employed for collecting reviews from Amazon website. Using this tool, a new dataset consisting of reviews from 27 hijacked products and 26 valid products was collected. The same technology was used to obtain crucial information pertaining to the title and description of each product, and the products were chosen to reflect the respective categories of products in order to avoid bias.

4.2 Preprocessing

After collecting the raw data, the unnecessary columns were removed from the dataset, and the non-English reviews were translated using `googletrans` library for python. This library also took care of a few grammar errors. Additionally, extra white spaces were removed, as well as special characters.

4.3 Annotating Data

The newly gathered data has undergone a semi-automatic labelling process. A Python script was implemented, which examined the review text and title for keywords and classified them into two categories, namely **related** (labelled with **"Y"**) and **unrelated** (labelled with **"N"**). For example, two of the keywords selected were "book" and "movie", since there were no books or movies products collected, but there were plenty of reviews about books and movies. This method managed to annotate less than half of the reviews collected. Subsequently, manual labelling was executed for the remaining data, followed by verification of the previously categorized data. Missing values in review titles, review descriptions, product titles, and product descriptions were filled with empty strings. After the annotation process was completed, the data consisted of 9486 reviews that were labelled with "Y", 9265 that were labelled with "N", and 473 that were removed as they did not possess sufficient information to be classified as related or unrelated. Before moving forward with tokenization, the necessary product information was concatenated into a new column `text_data`, and the labels from the `isRelated` columns were transferred to the new `labels` column. The newly created `text_data` column comprised of the columns `reviewTitle, reviewDescription, title`, (specifying the title of the product) and `description` (specifying the description of the product) combined. Finally, the `labels` column was converted to binary labels : 1 for "Y" and 0 for "N".

4.4 Tokenization

For the tokenization process, the RoBERTa tokenizer and BERT tokenizer are utilized respectively. The process of Bert tokenization involves the utilization of WordPiece tokenization, which divides words into subwords based on their frequency. This method is useful for handling words outside the vocabulary by breaking them into known subword elements. For example, the word "unhappiness" might be tokenized as `[un, ##happy, ##ness]`. The special tokens used by BERT are: `[CLS]` at the beginning of the input sequence, and `[SEP]` to separate sentences or mark the end of a sequence. For example, "`[CLS] This product is amazing![SEP]`". Contrarily, the RoBERTa tokenizer undergoes a tokenization procedure known as byte-pair-encoding (BPE). BPE is a technique employed to amalgamate the most common pairs of bytes within a text. It ensures that infrequent words are processed in a flexible manner, and is therefore very useful for large vocabulary. Furthermore, it uses `<s>` at the beginning and `</s>` at the end of each sentence, or to delimit different sentences. For example, "`<s> This product is amazing!</s>`" The RoBERTa Tokenizer also performs pre-processing steps like normalization and byte-level BPE.

4.5 Training and Evaluation

After preprocessing the dataset, it was divided into 10 subsets with balanced labels using Stratified K-fold Cross-Validation. For each fold, the subsequent operations were executed: tokenization was executed using the designated model's tokenizer, incorporating truncation and padding to achieve a maximum length of 256 tokens for BERT and 128 tokens for RoBERTa. Moreover, `DataLoader` objects were created from training and evaluation datasets to allow batch processing. Both the BERT and RoBERTa models were trained for a duration of three epochs per fold. The hyperparameters used for the models are depicted in Table 1 and Table 2 respectively.

5 Models

5.1 Model Architecture

The main focus of this research is to investigate whether RoBERTa model has a significant improvement in the accuracy of detecting hijacked reviews. Therefore, it is crucial to observe how RoBERTa performs on the collected data as opposed to BERT's performance. This chapter outlines the methodological framework used to answer the first research question.

5.1.1 BERT

BERT stands for Bidirectional Encoder Representations from Transformers. It is a model based on the transformer architecture, that revolutionized the field of natural language processing. A BERT architecture is made of multiple layers of bidirectional Transformer encoders. BERT base uses 12 encoders with 12 bidirectional self-attention heads, totalling 110 million parameters [26]. Input representation is represented by one token sequence. Therefore, a single sentence or multiple sentences can be represented by one token sequence. In our case, a sequence is made of review title, review description, title of the product and description of the product. Figure 21 shows in detail how multiple sentences are fed into the BERT architecture for text classification. BERT uses WordPiece embeddings with a 30,000 token vocabulary. The BERT model used in this study is `bert-base-uncased`, fine-tuned on our dataset for binary classification.

Hyperparameters	Values
batch_size	3
Epochs	3
Optimizer	AdamW
Loss Function	Cross-Entropy Loss
Learning rate	2^{e-5}
max_length sequence	256

Table 1: BERT's Hyperparameters used

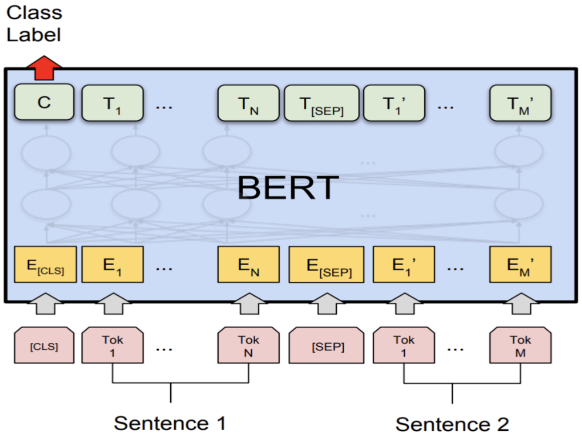


Figure 1: BERT architecture overview[1]

5.1.2 RoBERTa

RoBERTa (A Robustly Optimized BERT Pretraining Approach) is an optimized version of BERT [27]. It uses dynamic masking and is trained on more data for longer periods, which enhances its performance on downstream tasks. The RoBERTa model used is `roberta-base`, fine-tuned on our dataset as well.

Hyperparameters	Values
batch_size	8
Epochs	3
Optimizer	AdamW
Loss Function	Cross-Entropy Loss
Learning rate	2^{e-5}
max_length sequence	128

Table 2: RoBERTa's Hyperparameters used

5.2 Model Comparison

As previously mentioned, hyperparameters such as **`max_length`** and **`batch_size`** were determined based on the available GPU memory. As the length of the sequence increases, memory usage increases exponentially [37, 38]. A batch size of 32 and a maximum length of 38.5 GB required 11.5 gigabytes of memory and 38.5 GB of GPU to run the BERT classifier [37]. The machine on which the experiments were conducted utilized a GPU with a capacity of 6 GB. It is estimated that, using the same max_length sequence and batch_size for RoBERTa, training time would take approximately 30 hours. Hence, it was unfeasible to select the maximum values for **`maximum_length`**(512) and **`batch_size`**(32), and it was necessary to fine tune them according to the available resources. With these selected parameters, the training times for both models were equal. Both models are based on a common architecture, with modifications primarily affecting the training capabilities. Moreover, as the accuracy of the BERT model and RoBERTa model increases with increasing sequence length and batch size and our findings favour RoBERTa, it can be inferred that the results of the comparison are valid.

The training and validation loss are calculated and accumulated during each epoch's training loop, and the average training loss is printed after each epoch. The training accuracy is calculated by taking the average of the accuracy values for each batch of training data during training. To ensure consistency across the dataset and results, we used a stratified K-Fold Cross validation.

5.2.1 Evaluation Metrics

The following metrics with their respective formulae are calculated :

$$\text{True Positive Rate (TPR) or Recall} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{True Negative Rate (TNR)} = \frac{TN}{TN + FP} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN} \tag{4}$$

$$\text{False Negative Rate (FNR)} = \frac{FN}{FN + TP} \tag{5}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

[1] Source : `https://pytorch.org/tutorials/_images/bert.png`

Models Metrics	BERT	RoBERTa
Accuracy	0.9067	0.9302
F1	0.8118	0.8483
FPR	0.0628	0.0475
FNR	0.1815	0.1325
Recall	0.8600	0.8675
Precision	0.8022	0.8161
TNR	0.9372	0.9524
TPR	0.8601	0.8675

Table 3: Evaluation metrics results of both models for 10 folds

5.2.2 K-Fold Cross Validation

The models were validated by utilizing the Stratified K-fold Cross-Validation technique. Each model is trained multiple times, each time using k-1 folds for training and the remaining fold for testing. The procedure is repeated until each fold has been used as the test set precisely once. This procedure was performed for $k = 2, ..., 10$ as illustrated in Table 4. Stratified K-fold Cross-Validation ensures that each fold is representative of the entire dataset, thereby preserving the equilibrium between the two labels. To ensure there is no data leakage, the products have been separated to ensure that the models are trained on different products than the ones they are being tested on.The dataset was sorted by products, and it was verified that no product was present in the training and evaluation dataset at every fold.

Results (k=10)		
Folds	RoBERTa	BERT
2	0.9310	0.9085
3	0.9163	0.8988
4	0.9214	0.9123
5	0.9211	0.8841
6	0.9310	0.9031
7	0.9303	0.9104
8	0.9326	0.9066
9	0.9418	0.9338
10	0.9461	0.9028
μ	0.9302	0.9067
σ	0.00967	0.00017

Table 4: Stratified K-fold Validation Accuracy results of the two models

5.3 Error Analysis

Both models displayed low FPRs. This indicates effective minimization of incorrect classification of valid reviews as hijacked. The FPR and FNR of both models are depicted in Table 3. This suggests that RoBERTa was more conservative when flagging reviews as being hijacked. As illustrated in Table 2 and Table 1, AdamW optimizer was applied to both models. This optimizer incorporates a weight decay term to prevent overfitting and ensures that models are not overly tailored to training data. Their generalization is thus improved to unknown data.

To support, stabilize and avoid significant fluctuations in loss during the training process, the Learning Rate Scheduler was configured with warm-up steps. Moreover, training was conducted for three epochs, which included back propagation. Therefore, weight updates were based on weight loss calculated from training data, which resulted in a gradual reduction in errors.

5.3.1 Results

The Stratified K-Fold Cross-Validation presented in Table 4 indicates that RoBERTa exhibits a slight advantage over BERT. Additionally, considering the metrics used in Table 3, RoBERTa appears to have a slight advantage over BERT. However, we will focus on showing that there is a significant increase in accuracy for $\alpha = 0.05$. Accuracy is a relevant metric in this instance, as the dataset exhibits a balanced ratio of 50.59/49.41. We conducted a t-test to verify that this is indeed statistically significant for $\alpha = 0.05$.

5.3.2 T-Test Statistic

Using the samples from Table 4, we computed a Shapiro-Wilk test statistic to test for normality for $\alpha = 0.05$. Results are presented in Table 5. Therefore, we can assume normality for these two samples.

samples	BERT	RoBERTa
p-value	0.4063	0.6581

Table 5: Shapiro-Wilk test for normality for both samples, for $\alpha = 0.05$ using data from Table

We define the Null Hypothesis H_0 as follows: The mean accuracy of BERT is equal to the mean accuracy of RoBERTa.

$$H_0 : \mu_{\text{RoBERTa}} = \mu_{\text{BERT}}$$

Moreover, we define the Alternative Hypothesis H_1as follows: The mean accuracy of RoBERTa is greater than the mean accuracy of BERT.

$$H_1 : \mu_{\text{RoBERTa}} > \mu_{\text{BERT}}$$

The results are: **t-statistic**: 3.98 and **p-value**: 0.0013. We reject the null hypothesis test based on the p-value of 0.0013, which is significantly lower than the standard level of significance of 0.05. Hence, the evidence supports the assertion that the mean accuracy of the RoBERTa model surpasses that of the BERT model.

The analysis indicates that the RoBERTa model has a statistically significant increase in mean accuracy compared to the BERT model for the review hijacking across different cross-validations folds. Therefore, it appears that RoBERTa achieves superior performance when compared to BERT for $\alpha = 0.05$ in the detection of hijacked reviews when using real data collected from Amazon.

5.4 Determinants of Model Performance

There are several factors that influence the performance of these natural language processing models. Firstly, the training dataset might include some reviews that are categorized as **"related"** or **"unrelated"** even though this cannot be decided. This error is due to manual and semi-automatic labelling. There are roughly 5% of reviews that generally discuss the experience without explicitly mentioning the product or its usage, for example *"Great product"*, *"I love it"*, *"My daughter enjoys it very much"*. Despite the removal of numerous reviews, a significant number remains present in the training data, which can have an impact on the classification of reviews. Another factor that has an impact on the accuracy of the models is the loss of expressivity and nuance in reviews when translated from their original language. These processes are highly labor-intensive and require significant time investment. Consequently, it demands substantial resources, including manpower, time, and material assets.

6 Discussion

In this section, we discuss the potential factors that have been noticed to influence the possibility of a product being hijacked and how, based on these observations, we can support the design choices made for the visual tool concept.

6.1 Time distribution of reviews

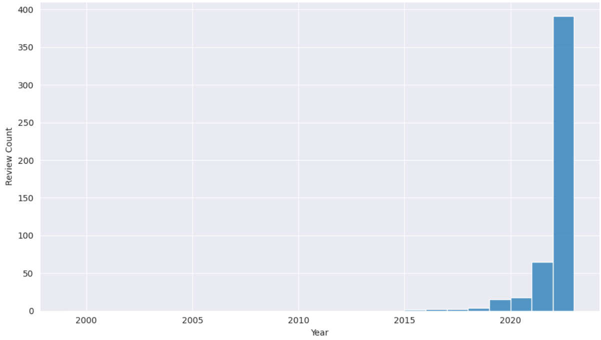


Figure 2: Review Count Distribution over Time for all Valid Products

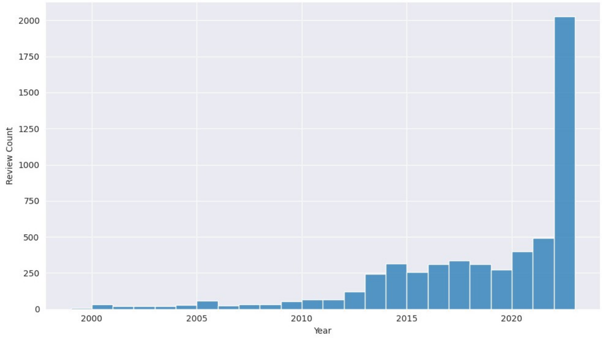


Figure 3: Review Count Distribution over Time for all Hijacked Products

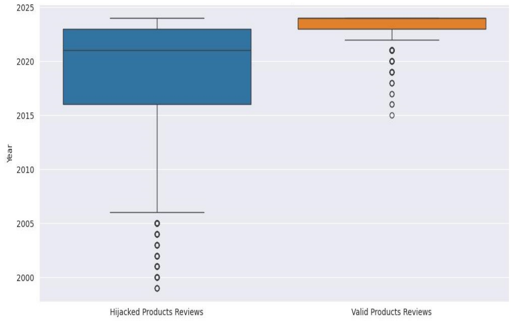


Figure 4: Box Plot Distribution of Review Years for Hijacked and Valid Products

The box plot depicted in Figure 4 presents a wide spectrum of years for hijacked products, with many reviews dating back to as early as 2000. The median is around 2020, but the lower quartile extends to earlier years, indicating the existence of many outdated reviews. In contrast, the range of valid product reviews is significantly restricted, with the majority of reviews centred around the year 2023. The median is also around 2023, indicating that valid products have more recent reviews.

Additionally, by visualizing the histogram, it supports the idea that there is a clear distinction in the distribution of review count over time for valid reviews as compared to hijacked product reviews. Figure 2 illustrates the review count distribution for valid reviews, while Figure 3 shows the corresponding distribution for hijacked reviews.

As illustrated in Figure 2, the volume of valid reviews is established over recent years, with a noticeable increase in recent years. In contrast, Figure 3 highlights a significant skew towards older reviews, indicating that most of the hijacked reviews are dated. It can be observed visually that the distribution of review count over time for hijacked products is more left-tailed. This distinction shall be utilized in the calculation of the credibility score.

6.2 Country of origin of reviews

During the collection of reviews for the dataset, it was observed that hijacked products had the majority of reviews from countries apart from the one the product was being sold in. This turned out to be a significant factor in identifying hijacked products.

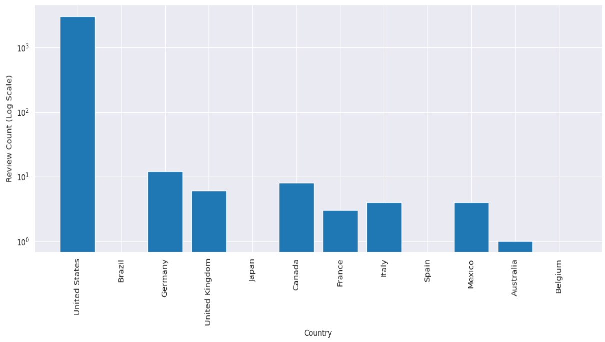


Figure 5: Review Count Distribution by Country for all Valid Products

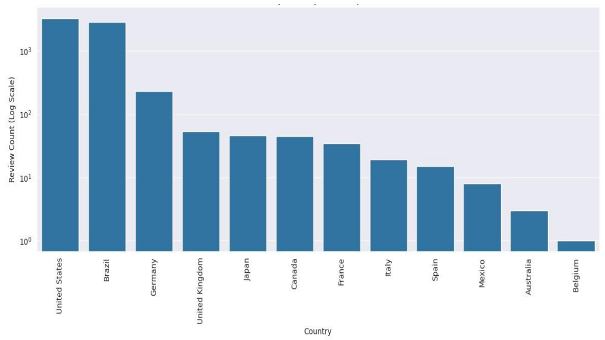


Figure 6: Review Count Distribution by Country for all Hijacked Products

Based on Figure 6 and Figure 5, it is clear that hijacked products exhibit a broader range of review counts across diverse nations. There is also a significant increase in reviews from other countries than the one the product is being sold in, namely the United States. We introduce the term *"country mismatch factor"(CMF)*, which measures the proportion of reviews that are not from the country the product is sold in.

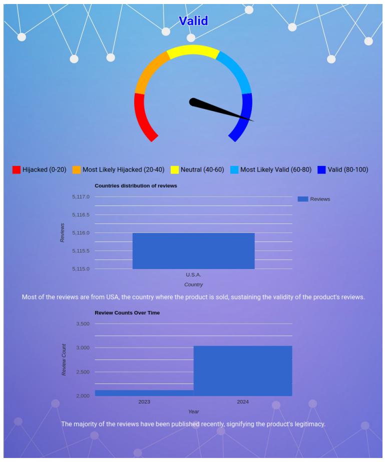
$$CMF = \frac{TotalReviews - TargetCountryReviews}{TotalReviews}$$

6.3 Seller Information

It has been observed that sellers who list hijacked products have few products and that most of them are hijacked. This is an important aspect in detecting hijacked products. Their products either have very few or no reviews, or they have thousands of reviews from reusing reviews. This aspect can be effectively utilized to alert the user about the potential hijacking of a product. However, this goes beyond this research, as there are plenty of tools online that assess sellers' credibility.

7 Visual Tool Concept

Based on all the important factors for review hijacking detection observed during this research, a visual tool concept was implemented. This visual tool is designed to efficiently communicate to the user the likelihood that the searched product has been hijacked. This is accomplished by utilizing data visualization techniques such as pie charts, histograms, and credibility scores by applying a formula that encompasses the factors observed.



Most of the reviews are from USA, the country where the product is sold, sustaining the validity of the product's reviews.

The majority of the reviews have been published recently, signifying the product's legitimacy.

Figure 7: Visual Tool Concept Valid Product With Visual Components

7.1 Credibility Score

Given the observations made earlier, the credibility score formula for the visualisation tool concept should incorporate these aspects to accurately communicate the validity of an online product. Therefore, the following formula is introduced:

$$CS = \gamma(1 - \frac{ValidReviews}{TotalReviews}) + \alpha(1 - K) + \beta(1 - CMF)$$
$$were\ \gamma = 0.8, \alpha = 0.15, \beta = 0.05,$$
$$and\ K\ is\ the\ kurtosis\ of\ Time\ Review\ Count\ Distribution$$

Each component of the formula bears its respective weight in determining the ultimate credibility score. γ was chosen to have the highest weight because the detection of related or unrelated reviews has the highest relevancy when detecting hijacked products. The components of the formula including kurtosis and country mismatch factor are indicative of review hijacking. However, there are instances where a hijacked product may not be detected if solely these two factors were to be considered. Therefore, these two terms have been chosen to have small weights.

7.2 Gauge Chart

The Gauge Chart was used to incorporate the choice of communicating the credibility score through colour coding. Colour coding is an effective way of communicating data [39, 40]. Raw data presented without any visual aid or summary can often fail to convey the intended message effectively [41]. Hence, the subsequent colour classification scheme was selected: red for **"Hijacked"**, orange for **"Most Likely Hijacked"**, yellow for **"Neutral"**, light blue for **"Most Likely Valid"**, and dark blue for **"Valid"**. Initially, green was chosen for determining the validity of a product; however, after considering potential users with colour blindness, it was determined that blue was a more suitable choice [42, 43, 39]. Each classification and its relationship to the numerical scale of the credibility score are explained under the gauge chart as depicted in Figure 7 and Figure 8.

7.3 Bar Chart

Initially, a pie chart was employed for the country distribution, despite the ambiguous scientific evidence regarding its utility [44, 45, 46]. It's harder to see when there are many subdivisions and when two or more of them are the same size. This may be partly because it is easier to measure length versus angle, both in terms of accuracy and visual judgment [47]. Therefore, the choice of a bar chart appeared superior [48]. A similar motivation goes for reviews count over time. A histogram is used to efficiently visualize how far back reviews go for a product, as well as their number. Both visual elements are accompanied by text that elucidates the observations made regarding these distributions and their significance in determining the product's validity.

8 Limitations

In this section, we will discuss each of the limitations of the research. The limitations of this study are computational and resource-related.

The dataset used for training and validation should be larger to achieve higher accuracy in the detection of hijacked products. The outcomes suggest that this can be accomplished if additional resources are utilized to collect more data, as previous research achieved up to 96% accuracy using 30M reviews obtained **synthetically**.

This study utilized a machine equipped with a GeForce RTX 2060 graphic card, with training times of up to 10 hours. Due to the exclusive utilization of this equipment, several hyperparameters had to be adjusted in accordance with the available memory for training purposes. If this limitation could be circumvented, different experiments would take place, as batch size could be increased, longer training times could be performed, as well as a bigger `max_length` sequence.

Since manual labelling was conducted, there is a possibility of potential errors being committed. The semi-automatic labelling process, however, has a higher potential for error. It is therefore recommended that extensive manual verification be conducted to minimize this as much as possible.

Due to the incompatibility of the `googletrans` library's with Google Translate, translation errors have been observed. There are several limitations of this tool, therefore more manual labouring should be employed to make up for these errors, as not all reviews were successfully translated or correctly translated.

9 Future Work

Future research should prioritize addressing the limitations of this study. It is imperative to gather a larger dataset and train models for extended periods of time. However, this requires the utilization of superior equipment to guarantee memory availability during the training process. Additional resources and manpower should be allocated to ensure that the manual labelling and translation errors are resolved.

Another topic for future research is the formula for calculating the credibility score. Future work should include the experimentation of different weights for the factors. Moreover, the observed factors that contribute to the detection of products that reuse reviews, such as the origin of reviews by country, the time distribution of reviews, and seller information, could be incorporated into the model to enhance its effectiveness.

The visual concept tool ought to be utilized by other researchers as a baseline for effectively communicating the legitimacy of a product in terms of review hijacking, and it can be enhanced to potentially become a web extension or web application. Furthermore, it could be incorporated into the pre-existing web extensions that evaluate the seller's credibility and potential fake reviews, as the seller's credibility is a crucial aspect of review hijacking as well.

10 Conclusion

During the course of this investigation, two language processing models were evaluated on a dataset consisting of both hijacked and legitimate product reviews obtained from Amazon. We conclude that both BERT and RoBERTa are suitable tools for identifying hijacked reviews. However, upon analysing the outcomes of both models, done in subsections 5.3.1 and 5.3.2, RoBERTa exhibits a slightly superior performance in our experiments, which answers (RQ_1). We propose a formula for calculating the **credibility score** that incorporates the observations made during this research. The subject of credibility score and its elements are addressed in subsections 6.1, 6.2 and 7.1 and provides support for (RQ_2). Furthermore, we justify the choice of each element for the visual concept tool to communicate the credibility score effectively to potential users. Hence, we arrive at the conclusion that a credibility score, founded on the model's ability to detect hijacked reviews, which reflects the likelihood that a product is hijacked, can be effectively communicated to potential users through visual elements accompanied by text that provides additional elucidation. This subject is covered in section 7 and answers the last research question (RQ_3). In order to provide future research with beneficial tools for effectively combating review hijacking, the limitations of this study are also presented along with ways to overcome them. To summarize, this research provides a dataset to the field of Natural Language Processing (NLP) in e-commerce, enabling the development and evaluation of techniques for identifying review hijacking.

References

[1] Dina Mayzlin. Promotional Chat on the Internet. *Marketing Science*, 25(2):155–163, 03-04 2006.

[2] Sung Ha, Soon Yong Bae, and Lee Son. Impact of online consumer reviews on product sales: Quantitative analysis of the source effect. *Applied Mathematics Information Sciences An International Journal*, 9:373, 2015.

[3] Judith A. Chevalier and Dina Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354, 2006.

[4] Mengzhou Zhuang, Geng Cui, and Ling Peng. Manufactured opinions: The effect of manipulating online product reviews. *Journal of Business Research*, 87:24–35, 2018.

[5] Kendall L Short. Buy my vote: online reviews for sale. *Vand. J. Ent. & Tech. L.*, 15:441, 2012.

[6] Yubo Chen and Jinhong Xie. Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science*, 54:477–491, 03 2008.

[7] Sonja Utz, Peter Kerkhof, and Joost van den Bos. Consumers rule: How consumer reviews influence perceived trustworthiness of online stores. *Electronic Commerce Research and Applications*, 11(1):49–58, 2012. Special Issue: CAT Tournament.

[8] Sammy Paget. Local consumer review survey: How customer reviews affect behavior, 02 2023. `https://www.brightlocal.com/research/local-consumer-review-survey`, (visited 05 2024).

[9] Kaitlin A. Dohse. Fabricating feedback: Blurring the line between brand management and bogus reviews. *University of Illinois Journal of Law, Technology Policy*, 2013:364, 2013.

[10] Ling Peng, Geng Cui, Mengzhou Zhuang, and Chunyu Li. Consumer perceptions of online review deceptions: an empirical study in china. *Journal of Consumer Marketing*, 33(4):269–280, 2016.

[11] Arjun Mukherjee, Bing Liu, and Natalie Glance. Spotting fake reviewer groups in consumer reviews. *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web*, pages 191–200, 05 2012.

[12] Nafiz Sadman, Kishor Datta Gupta, Ariful Haque, Subash Poudyal, and Sajib Sen. Detect review manipulation by leveraging reviewer historical stylometrics in amazon, yelp, facebook and google reviews. In *Proceedings of the 2020 The 6th International Conference on E-Business and Applications*, ICEBA 2020, page 42–47, New York, NY, USA, 2020. Association for Computing Machinery.

[13] Timothy B. Lee. Amazon needs to crack down on bait-and-switch listings, 12 2021. `https://slate.com/technology/2021/12/amazon-listings-wrong-reviews-why.html`, (visited 05 2024).

[14] Monika Daryani and James Caverlee. Identifying hijacked reviews. In Shervin Malmasi, Surya Kallumadi, Nicola Ueffing, Oleg Rokhlenko, Eugene Agichtein, and Ido Guy, editors, *Proceedings of the 4th Workshop on e-Commerce and NLP*, pages 70–78, Online, August 2021. Association for Computational Linguistics.

[15] Aishwarya Deep Shukla and Jie Mein Goh. Fighting fake reviews: Authenticated anonymous reviews using identity verification. *Business Horizons*, 67(1):71–81, 2024.

[16] Nicole Nguyen. Here's another kind of review fraud happening on amazon, 05 2018. `https://www.buzzfeednews.com/article/nicolenguyen/amazon-review-reuse-fraud`, visited 05 2024).

[17] J. Swearingen. Hijacked reviews on amazon can trick shoppers. https://www.consumerreports.org/customer-reviews-ratings/hijacked-reviews-on-amazon-can-trick-shoppers/, aug 26 2019.

[18] Huma Qayyum, Farooq Ali, Marriam Nawaz, and Tahira Nazir. Frd-lstm: a novel technique for fake reviews detection using dcwr with the bi-lstm method. *Multimedia Tools and Applications*, 82, 03 2023.

[19] Juan María Martínez Otero. Fake reviews on online platforms: perspectives from the us, uk and eu legislations. *SN Social Sciences*, 1, 07 2021.

[20] Gourav Bathla, Pardeep Singh, Rahul Kumar Singh, Erik Cambria, and Rajeev Tiwari. Intelligent fake reviews detection based on aspect extraction and analysis using deep learning. *Neural Computing and Applications*, 34, 07 2022.

[21] Ahmed Elmogy, Usman Tariq, Ammar Mohammed, and Atef Ibrahim. Fake reviews detection using supervised machine learning. *International Journal of Advanced Computer Science and Applications*, 12, 01 2021.

[22] Sherry He, Brett Hollenbeck, and Davide Proserpio. The market for fake reviews. *Marketing Science*, 41(5):896–921, 2022.

[23] Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon gyo Jung, and Bernard J. Jansen. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64:102771, 2022.

[24] Alexandre Yukio Ichida, Felipe Meneguzzi, and Duncan D. Ruiz. Measuring semantic similarity between sentences using a siamese neural network. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2018.

[25] Bharath Chandra Chikoti. Text similarity using siamese networks and transformers. *International Journal for Research in Applied Science and Engineering Technology*, 10:1856–1863, 06 2022.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[28] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, page 219–230, New York, NY, USA, 2008. Association for Computing Machinery.

[29] Somayeh Shojaee, Masrah Azrifah Azmi Murad, Azreen Bin Azman, Nurfadhlina Mohd Sharef, and Samaneh Nadali. Detecting deceptive reviews using lexical and syntactic features. In *2013 13th International Conference on Intellient Systems Design and Applications*, pages 53–58, 2013.

[30] Priyanka Gupta, Shriya Gandhi, and Bharathi Raja Chakravarthi. Leveraging transfer learning techniques- bert, roberta, albert and distilbert for fake review detection. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '21, page 75–82, New York, NY, USA, 2022. Association for Computing Machinery.

[31] Huma Ayub, Farooq Ali, Marriam Nawaz, and Tahira Nazir. Frd-lstm: a novel technique for fake reviews detection using dcwr with the bi-lstm method. *Multimedia Tools and Applications*, 82:1–15, 03 2023.

[32] Dong Zhang, Wenwen Li, Baozhuang Niu, and Chong Wu. A deep learning approach for detecting fake reviewers: Exploiting reviewing behavior and textual information. *Decision Support Systems*, 166:113911, 2023.

[33] Rami Mohawesh, Haythem Bany Salameh, Yaser Jararweh, Mohannad Alkhalaileh, and Sumbal Maqsood. Fake review detection using transformer-based enhanced lstm and roberta. *International Journal of Cognitive Computing in Engineering*, 5:250–258, 2024.

[34] Rami Mohawesh, Shuxiang Xu, Son N. Tran, Robert Ollington, Matthew Springer, Yaser Jararweh, and Sumbal Maqsood. Fake reviews detection: A survey. *IEEE Access*, 9:65771–65802, 2021.

[35] Ana Reyes-Menendez, José Saura, and Ferrao Filipe. The importance of behavioral data to identify online fake reviews for tourism businesses: A systematic review. *PeerJ Computer Science*, 5, 09 2019.

[36] Petr Hájek and Jean-Michel Sahut. Mining behavioural and sentiment-dependent linguistic patterns from restaurant reviews for fake review detection. *Technological Forecasting and Social Change*, 177:121532, 04 2022.

[37] Ali Areshey and Hassan Mathkour. Transfer learning for sentiment classification using bidirectional encoder representations from transformers (bert) model. *Sensors*, 23(11), 2023.

[38] Christian Y. Sy, Lany L. Maceda, Mary Joy P. Canon, and Nancy M. Flores. Beyond bert: Exploring the efficacy of roberta and albert in supervised multiclass text classification. *International Journal of Advanced Computer Science and Applications*, 15(3), 2024.

[39] Steven L. Franconeri, Lace M. Padilla, Priti Shah, Jeffrey M. Zacks, and Jessica Hullman. The science of visual data communication: What works. *Psychological Science in the Public Interest*, 22(3):110–161, 2021. PMID: 34907835.

[40] Markus Christen, Peter Brugger, and Sara Irina Fabrikant. Susceptibility of domain experts to color manipulation indicate a need for design principles in data visualization. *Plos one*, 16(2):e0246479, 2021.

[41] Anne Helmond. 'raw data' is an oxymoron. *Information*, 17, 10 2014.

[42] Fabio Crameri, Grace E Shephard, and Philip J Heron. The misuse of colour in science communication. *Nature communications*, 11(1):5444, 2020.

[43] Andrew Disney. Choosing colors for your data visualization, 11 2020. `https://cambridge-intelligence.com/choosing-colors-for-your-data-visualization/`, (visited 05 2024).

[44] Ian Spence. No humble pie: The origins and usage of a statistical chart. *Journal of Educational and Behavioral Statistics*, 30(4):353–368, 2005.

[45] Ian Spence and Stephan Lewandowsky. Displaying proportions and percentages. *Applied Cognitive Psychology*, 5(1):61–77, 1991.

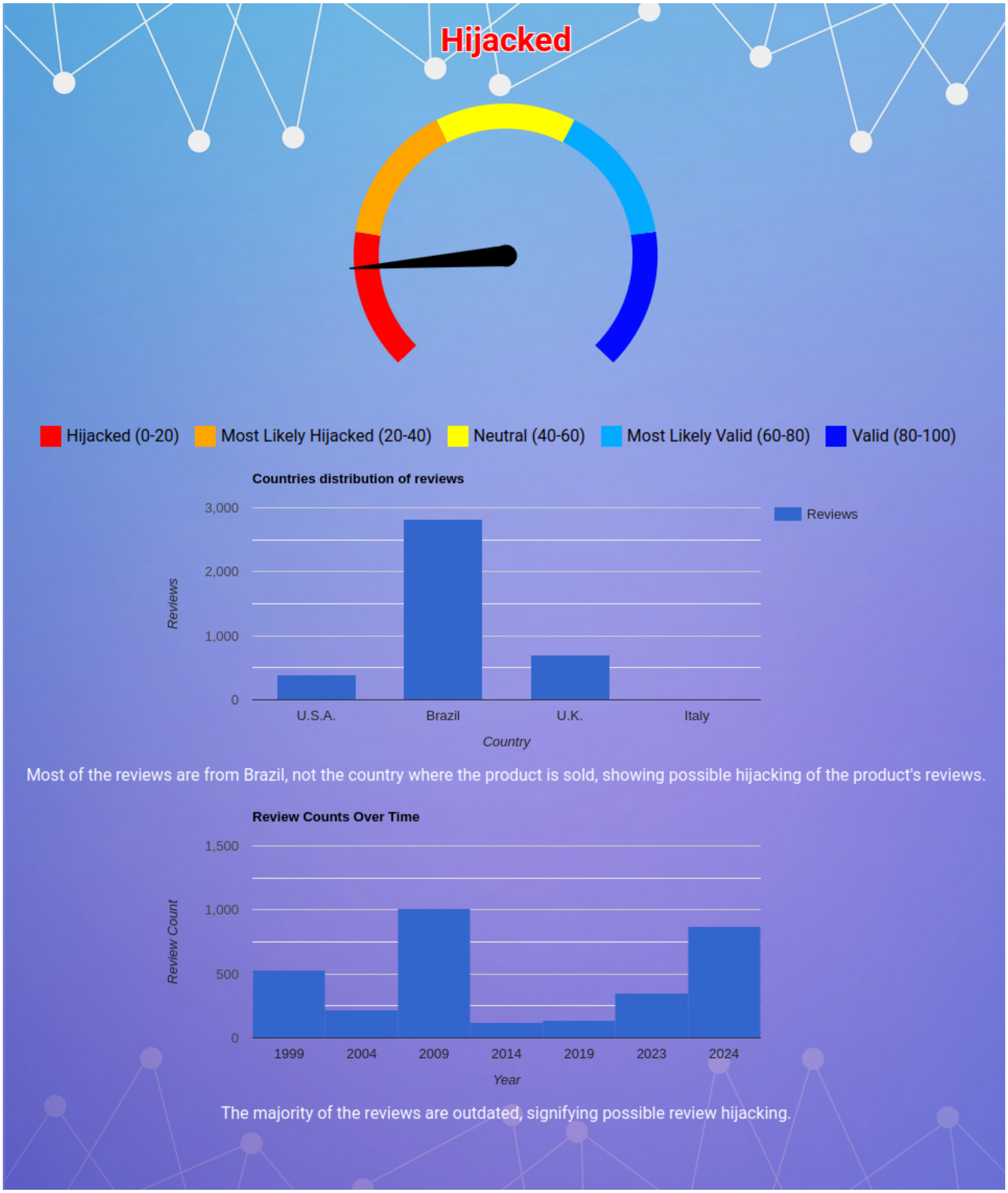
[46] Deb Feldman-Stewart, Nancy Kocovski, Beth A McConnell, Michael D Brundage, and William J Mackillop. Perception of quantitative information for treatment decisions. *Medical Decision Making*, 20(2):228–238, 2000.

[47] Chris Pritchard. Life of pie: William playfair and the impact of the visual, 07 2021. `https://www.m-a.org.uk/resources/PE4LifeofPie.pdf`, (visited 05 2024).

[48] World Health Organization et al. Tools for making good data visualizations: the art of charting. Technical report, World Health Organization. Regional Office for Europe, 2021.

A Appendix

A.1 Visual Tool



Most of the reviews are from Brazil, not the country where the product is sold, showing possible hijacking of the product's reviews.

The majority of the reviews are outdated, signifying possible review hijacking.

Figure 8: Visual Tool Concept Hijacked Product With Visual Components