Revista Brasileira de Computação Aplicada, April, 2025

DOI: 10.5335/rbca.v17i1.16183
Vol. 17, Nº 1, pp. 12–22
Homepage: seer.upf.br/index.php/rbca/index

ORIGINAL PAPER

Benchmarking Machine Learning Algorithms in Fake Reviews Detection in Brazilian Portuguese

Eduardo Camargo Ribeiro Borges [1], Cristiano Mesquita Garcia [1], Samuel da Silva Feitosa [2], Carlos Henrique Radavelli [1]

[1]Instituto Federal de Santa Catarina (IFSC), [2]Universidade Federal da Fronteira Sul (UFFS)

* ecamargoborges@gmail.com; cristiano.garcia@ifsc.edu.br; samuel.feitosa@uffs.edu.br; carlos.radavelli@ifsc.edu.br

Received: 2024-08-19. **Revised:** 2025-04-25. **Accepted:** 2025-04-30.

Abstract

The proliferation of fake reviews has become a growing concern on e-commerce platforms, as these reviews can mislead consumers and harm the reputation of products and services offered. Automatic detection of fake reviews is a challenging task, as it requires analyzing textual data and identifying subtle patterns that indicate the veracity of reviews. Since fake review datasets in Portuguese are scarce, in this work, we generate and propose a dataset in Brazilian Portuguese for the detection of fake reviews. Then, four machine learning algorithms, combined with three text vectorization methods, are used in a transfer learning scheme for fake review classification. A comparative analysis is carried out using performance metrics such as accuracy, F1-score, and false positives. The results show that, for the proposed dataset, the combination of Logistic Regression and a pre-trained BERT model in Brazilian Portuguese, i.e., BERTimbau, reached the best metric values, reaching 96.61% of accuracy.

Keywords: Fake reviews; Machine learning; Classification; Natural Language Processing.

Resumo

A proliferação de avaliações falsas tornou-se uma preocupação crescente nas plataformas de comércio eletrônico, uma vez que essas avaliações podem induzir os consumidores ao erro e prejudicar a reputação dos produtos e serviços oferecidos. A detecção automática de avaliações falsas é uma tarefa desafiadora, pois exige a análise de dados textuais e a identificação de padrões sutis que indicam a veracidade das avaliações. Dado que conjuntos de dados de avaliações falsas em português são escassos, este estudo gerou e propôs um conjunto de dados em português brasileiro para a detecção de avaliações falsas. Foram utilizados quatro algoritmos de aprendizado de máquina, combinados com três métodos de vetorização de texto, em um esquema de aprendizado por transferência para a classificação de avaliações falsas. Foi realizada uma análise comparativa utilizando acurácia, F1-score e falsos positivos. Os resultados mostram que, para o conjunto de dados proposto, a combinação de Regressão Logística e um modelo BERT pré-treinado em português brasileiro, i.e., BERTimbau, alcançou os melhores valores de acurácia, atingindo 96,61%.

Palavras-Chave: Avaliações falsas; Aprendizado de máquina; Classificação; Processamento de Linguagem Natural.

1 Introduction

E-commerce has been growing significantly in the last decades, boosted by the increase in Internet access and the availability of online transactions. However, implementing e-commerce was initially complex because customers did not trust this system (de Melo Cruz, 2021). With the popularization of the Internet, the situation changed, and customers have several options for marketplaces and e-commerce platforms. In addition, they can check reviews, provide product feedback for other users, and track their products through logistic systems. It is also noticeable that the Covid-19 pandemic accelerated companies' migration to online (da Costa et al., 2021),

leading to historical records in sales in Brazil (Ebit, 2021).

There are risks associated with online shopping, such as falsifications and scams. With the increase in online shopping, there is an increase in virtual scams. Sellers might use techniques to achieve competitive advantages and raise profits. A number of these companies have appealed to illicit practices, such as using fake reviews, to improve their reputation and increase sales (Cao, 2023).

This scenario led customers to be careful and prioritize trustworthy platforms for online shopping (Mota, 2021). Accounting for that, since customers also seek feedback from others when using e-commerce, they generally are confident that those platforms would hold reliable information. The use of Machine Learning (ML) can be an effective alternative to deal with the problem of fake reviews on e-commerce platforms (Mohawesh et al., 2021b).

Fake reviews are a relevant problem in e-commerce, since they are critical in purchasing process. According to BrightLocal, 82% of the customers had observed a fake review during a purchase, and citing a Washington Post research[1], 61% of the Amazon reviews are believed to be fake. In addition, Akesson et al. (2023) estimate the losses of USD 0.12 due to fake reviews per dollar spent. To demonstrate the impact of fake reviews, a reporter created a fake restaurant and made it reach the top 1 in London based on fake reviews[2].

Therefore, initiatives to detect and prevent fake reviews are welcome to avoid financial losses – for customers and companies –, and distrust by the customer, promoting a fair competition between e-sellers. Thus, this paper aims to evaluate the efficiency of machine learning algorithms in detecting fake reviews in Brazilian Portuguese. To the best of our knowledge, this is the first time a dataset in Portuguese has been proposed for detecting fake reviews.

Our contributions are 3-fold: (a) a dataset in Brazilian Portuguese to be used as a benchmark for fake reviews detection in such language; (b) a GPT-2 model in Brazilian Portuguese for fake review generation; and (c) a comparative analysis of traditional classifiers in the task of fake review detection.

This paper is organized as follows. Section 2 presents fundamental concepts; Section 3 describes the methodology of this work, Section 4 details the experimental protocol, Section 5 presents and discusses the obtained results, and finally, Section 6 concludes this paper.

2 Background

This section presents the fundamentals of the concepts applied in this paper: e-commerce and associated risks, fake reviews, and fake review detection.

2.1 E-commerce and its inherent risks

E-commerce has been growing significantly in the last decades. Access to the Internet and popularizing online transactions played an important role. The convenience of purchasing from home increased the preference for e-commerce platforms compared to traditional shopping (de Andrade and Silva, 2017). Notably, the Covid-19 pandemic accelerated physical stores' migration to digital or at least their online presence (da Costa et al., 2021).

The commercial use of the Internet began in the early 90s. However, it took a few years to be considered a business opportunity. Novaes (2016) mention that, from the 90s, the interest in business on the Internet started to be intensified, raising its commercial stage. According to de Melo Cruz (2021), in 1995, there already were online companies that later became huge market players, such as eBay and Amazon.

According to de Melo Cruz (2021), the beginning of e-commerce was difficult since customers did not trust the online shopping system. Nowadays, several options are available, easing the customer to check reviews and scores provided by previous customers. de Melo Cruz (2021) also mentions that the popularization of e-commerce correlates with access to the Internet in several countries.

With the Internet becoming popular, da Costa et al. (2021) mentions that, in 2000s, companies such as Americanas.com and Mercado Livre reached a great share of the Brazilian market. However, to da Costa et al. (2021), the penetration of e-commerce in Brazil was smaller (5.18%) than in countries such as China (approximately 28%). Therefore, even with the popularization of the Internet, the Brazilian market was not as representative.

The scenario changed at the beginning of 2020 due to the Covid-19 pandemic (da Costa et al., 2021). Companies with only physical stores had to be present in the digital world for survival. Thus, with the pandemics and the measures to avoid agglomerations, online shopping has become essential to purchasing and receiving products at home with almost no human contact. According to Ebit (2021), a steady growth in e-commerce sales is observed until 2019. However, in the first quarter (Q1) of 2020, the growth was 46% compared to the same period in 2019, and in 2021 Q1, online sales reached Brazilian real (BRL) 53.4 billion, corresponding to a growth of 31% in comparison with 2020 Q1.

According to de Andrade and Silva (2017), risks in online shopping are elevated, with falsifications and scams the most frequent risks. Mota (2021) stated that the growth in online shopping correlated with the increase in virtual scams. Therefore, e-customers tend to be careful and seek trustworthy e-commerce platforms to make purchases. Furthermore, when using e-commerce, customers have more time and can purchase whenever they want (Nascimento, 2011). According to de Andrade and Silva (2017), 57% of the surveyed people stated that the lack of security impedes the development of e-commerce in Brazil. To emphasize, recent news reported Brazilian reais (BRL) 2.5 billion in fraud attempts in the first half of 2023[3]

[1] https://www.washingtonpost.com/business/economy/how-merchants-secretly-use-facebook-to-flood-amazon-with-fake-reviews/2018/04/23/5dad1e30-4392-11e8-8569-26fda6b404c7_story.html?

[2] Available at: https://www.youtube.com/watch?v=bqPARIKHbN8. Accessed on March 12th, 2024.

[3] Available at: https://www.cnnbrasil.com.br/economia/brasil-reg

2.2 Fake Reviews

When selecting a product/service to purchase online, customers generally evaluated quantitative metrics, such as rating (commonly using a star scale between 1 and 5), and qualitative information, such as comments/reviews. Resorting to reviews before purchasing is becoming a frequent habit among customers (Cao, 2023). It is not a recent observation, since Valant (2015) mentioned a survey applied by the European Consumer Centres' Network obtained the information that 82% of the surveyed people read reviews on a product before purchasing.

Akesson et al. (2023) analyzed the impact of (fake) reviews. In experiments executed in 2020 with 1,000 adults from the UK, the authors concluded that inflated star ratings increase the chance of customers purchasing a "dont-buy" product. This chance increases with a fake review, reducing the chance of purchasing a "best-buy" product. The definitions of "dont-buy" and "best-buy" were performed by Which? (Which?, 2020), having the "dont-buy" as poor-quality products, and "best-buy" as good-quality products.

In this scenario, e-commerce companies leverage different ways to increase sales. Some may use unethical methods to prevail against competitors. Fake reviews are among the frequent methods. Cao (2023) state that e-sellers manipulate comments on their products to lure more buyers. These manipulated comments are fake reviews. Tufail et al. (2022) mention that fake reviews are created to influence products'/services' reputation in e-commerce platforms. According to Mohawesh et al. (2021b), these fraudulent reviews can enhance or poor a product/service/business' reputation. Mohawesh et al. (2021b) mention that positive reviews can lead to greater financial gains, while negative ones can cause losses to e-sellers. In addition, fake reviews can be produced either by people or machines. According to BrightLocal, 82% of the surveyed users had already observed a fake review, while around 61% of the reviews on Amazon are estimated to be fake.

Strategies for fake reviews may differ. In genuine customer cases, the e-seller may offer benefits to influence the feedback. According to Cao (2023), these benefits include cashback and voucher discounts. These benefits can influence the customer to write a review better than his/her opinion in positive cases, while being less rigorous in case of a negative review (Cao, 2023). According to Mohawesh et al. (2021b), this can also be considered a fake review.

Besides offering advantages to real customers, e-sellers also hire people to write reviews. He et al. (2022) described the fake review market. The authors mention that fake reviewers are recruited through social media platforms, such as Facebook. The Facebook groups were active, with some having around 16 thousand members. The researchers also concluded that e-sellers used these groups to promote their products, requesting the members to purchase and write a positive review in exchange for a total money return. In some cases, the more realistic the review, the more money the fake reviewer receives. He et al. (2022) also mention that, financially, this may be worth it for e-sellers since a small sales number may return the "investment".

2.3 Fake Review detection

Detecting fake reviews is not a trivial task, since the writing patterns of fake reviewers tend to change to surpass a possible fake review detection filter (Mohawesh et al., 2021a). However, there is the possibility that customers may detect fake reviews. This approach eases understanding fake reviewers' methods and patterns, indicating potential detection rules (Salminen et al., 2022). The problem with this technique is that once the fake reviewers understand the detection rules, their methods change to overcome the detection (Salminen et al., 2022; Mohawesh et al., 2021a).

Another expected aspect complicating the detection scenario is that fake reviews can be similar to genuine reviews (Ott et al., 2011). Therefore, machine learning methods for detection can be of great help. Ott et al. (2011) recruited people and used automated classifiers to evaluate reviews as fake or genuine; classifiers obtained the best performance in most metrics. Mohawesh et al. (2021b) mentioned that most researchers, using manual detection, obtained at most 60% of accuracy. Thus, using machine learning with the help of natural language processing for fake review detection constitutes an interesting alternative. In this case, the most suitable learning process is supervised learning, in which the researchers have examples of fake and genuine reviews (Mohawesh et al., 2021b).

3 Methodology

This paper utilizes the methodology proposed by Salminen et al. (2022), which generates fake reviews using a fine-tuned Generative Pre-trained Transformer (GPT), version 2 (GPT-2)(Topal et al., 2021; Solaiman et al., 2019; Radford et al., 2019).

More specifically, from a particular Amazon dataset, Salminen et al. (2022) identified the 10 most frequent categories. From these categories, the authors selected the one with the least frequency as a reference, leading to a selection of approximately 2,500 items per category, totaling 40,000 items for fine-tuning. After fine-tuning, Salminen et al. (2022) generated 2,000 reviews per category, using the first 5 words from sampled reviews as input for GPT-2, which should generate the remainder of the review. Therefore, 20,000 fake reviews were generated. In the end, Salminen et al. (2022) condensed this information into a balanced 40,000 dataset.

4 Experimental Protocol

This section details the experimental protocol applied in this paper, the description of the original dataset, the classifiers used for fake review detection, the text vectorization methods, the classifiers, the performance

istra-r-25-bi-em-tentativas-de-fraude-no-e-commerce-no-1o-semestre-diz-levantamento/. Accessed on March 12th, 2024.

metrics to evaluate the classifiers, and the evaluation strategies.

4.1 Dataset

This paper leverages the Brazilian E-Commerce Public Dataset by Olist(Olist and Sionek, 2018)[4]. This dataset consists of eight subsets, of which one is related to reviews. The reviews are linked to the orders. However, we see in Fig. 1 that the vast majority are single-product sales. In addition, the subset related to product description also had to be used to extract product category and perform similarly as Salminen et al. (2022). Thus, since most orders include only one item, we used the orders' reviews as if they were product reviews to keep in the methodology presented by Salminen et al. (2022). From the 100,000 records available, only 40,000 were considered useful since the review field was filled.

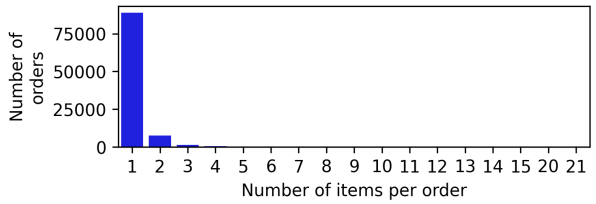

Figure 1: Items per order.

Following the methodology proposed by Salminen et al. (2022), we observed that the Olist dataset has fewer reviews than the dataset used by the authors. We had to select the 15 most frequent categories to have a final dataset of reasonable size, as shown in Fig. 2.

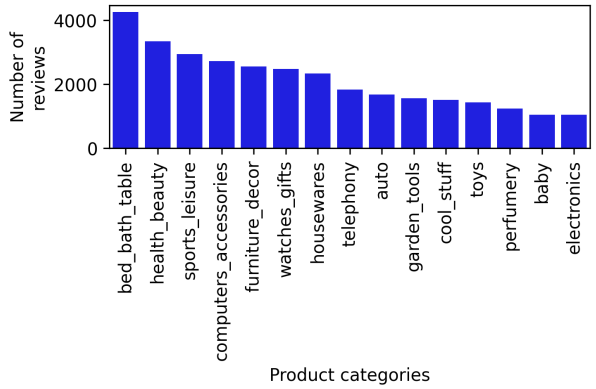

Figure 2: Category distribution, considering the 15 most frequent categories.

Therefore, we generated 1,000 fake reviews for each category using a fine-tuned GPT-2 for Brazilian Portuguese. Still, in line with (Salminen et al., 2022), we kept a balanced number of reviews per category, totaling 30,000 reviews. The GPT-2 model is available at https://huggingface.co/Ecrb/gpt2-reviews-ptbr.

4.2 Text Vectorization Methods

We applied Bag-of-Words (BoW) (Harris, 1954) and Term Frequency - Inverse of Document Frequency (TF-IDF) (Salton and Buckley, 1988) as text vectorization methods. BoW counts the occurrence of words using a tabular structure in which the features (or columns) are the words, and each row corresponds to a document (or text).

TF-IDF is similar to BoW (Garcia et al., 2025b), but instead of maintaining the counts, TF-IDF calculates the importance of words across documents using Eqs. (1) to (3), where d is a document or text.

$$TF(term, d) = \frac{count(term)}{sum(count(\text{terms in d}))}, \quad (1)$$

$$IDF(term) = log\frac{\text{number of documents}}{document\_frequency(term)} \quad (2)$$

$$TFIDF(term, d) = TF(term, d) \times IDF(term) \quad (3)$$

Furthermore, we included pre-trained representations in this paper. We used BERTimbau (Souza et al., 2020), a pre-trained BERT-based (Devlin et al., 2019) model in Brazilian Portuguese. Bidirectional Encoder Representation from Transformers (BERT) is a Transformer-based language model developed by Google (Devlin et al., 2019). It is extensively used in tasks such as classification (Thuma et al., 2023), spam detection (Otieno et al., 2023), etc. BERT is pre-trained using unlabeled data for masked language modeling (MLM) and next sentence prediction (NSP). In the MLM task, the BERT model learns the relationships between words in a sentence, while in the NSP, the relationships between sentences are learned (Devlin et al., 2019). These tasks increase BERT's capability of generalization and contextualization.

4.3 Classifiers

In this paper, we evaluate four classifiers: (a) Logistic Regression, (b) Decision Tree, (c) Random Forest, and (d) Support Vector Machine (SVM). We used Python 3.8 and Sci-kit Learn (Pedregosa et al., 2011), and the classifiers used default parameters.

4.3.1 Logistic Regression

Logistic Regression (LR) is a classification model suitable for binary problems (Cox, 1958; Nick and Campbell, 2007). Based on input, this model can calculate the probability of a binary event by applying a logistic function, ensuring

[4]Available at: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce. Accessed on March 13th, 2024.

a probabilistic interpretation (da Cruz Machado Benatti, 2017).

The relationship between the binary dependent and independent variables is quantified by estimating coefficients, typically using maximum likelihood estimation. Logistic regression is extensively utilized across various disciplines, including medicine, social sciences, and machine learning, for applications such as heart disease prediction (Godoy et al., 2023), credit risk assessment (Runchi et al., 2023), and spam detection (Berrou et al., 2023).

4.3.2 *Decision Tree*

Decision Tree (DT) is a supervised learning algorithm that learns data splits, representing the splits in a tree with split and leaf nodes. Each node represents a choice between (or among) different paths, generated according to criteria such as Gini index (de Almeida Teodoro and Kappel, 2020).

The tree is constructed by recursively partitioning the data set into subsets based on the attribute that yields the highest information gain or lowest impurity. Decision trees are valued for their interpretability and simplicity (de Almeida Teodoro and Kappel, 2020).

4.3.3 *Random Forest*

A Random Forest (RF) is an ensemble learning method that combines many Decision Trees, which generally uses different data partitions and features (Leite et al., 2023). The method introduces randomness by bootstrapping samples and considering a random subset of features for splitting at each node, enhancing model robustness and generalization. RFs can improve accuracy and reduce overfitting since they consider the votes of all trees when making a prediction (Leite et al., 2023).

4.3.4 *Support Vector Machine*

Support Vector Machine is a model that learns boundaries and maximizes the margins to create better separations between classes (da Cruz Machado Benatti, 2017). SVMs can achieve interesting performance when using high-dimensional features as input, which is the case with texts in general (Thuma et al., 2023).

4.4 Metrics

Since this paper generates a balanced dataset for detection, accuracy can be considered a suitable metric. This metric considers true positives (TP), true negatives (TN), false positives (FP), and false negatives (FP). Eq. (4) shows the formula for accuracy. In other words, accuracy corresponds to the ratio between hits and the number of classified items.

$$accuracy = \frac{TP + FN}{TP + TN + FP + FN} \tag{4}$$

In addition, we used the confusion matrix, which visually emphasizes the true and false components described above, and Macro F1-Score, which is insensitive to data imbalance. F1-Score (Eq. (7)) is based on precision (Eq. (5)) and recall (Eq. (6)). The rationale behind the F1-Score is that it requires both precision and recall to be high to achieve a high F1-Score.

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

F1-Score of a class is mathematically described in Eq. (7) as:

$$F1(class) = 2 \times \frac{precision_{class} \times recall_{class}}{precision_{class} + recall_{class}} \tag{7}$$

In all the above-mentioned metrics, the interpretation is that the closer to 1 (or 100%), the better. The F1-Scores are reported using the average of the classes.

4.5 Evaluation strategies

In this paper, we performed two evaluations. The first was a cross-validation performed considering a stratified K-fold strategy using 10 folds. Therefore, the results are averaged in terms of Accuracy and F1. This strategy is robust for model selection (Prusty et al., 2022).

Since we also report the results in terms of confusion matrices, we performed a hold-out strategy, splitting the data into 70% for training and 30% for testing in a stratified manner. The confusion matrices show the results obtained using the test set.

5 Results

This section presents the results, organized by text vectorization methods. First, it is important to state what the errors, i.e., false positive and false negative, mean. In our problem, a false positive means a genuine review is classified as a fake review. On the other hand, a false negative is a fake review classified as a genuine review. For analysis purposes, we assume false negatives are more prejudicial since the fake review would still be available to the user, who might be influenced by it to make bad purchase decisions.

The results are presented in terms of confusion matrices, while the metric values are reported in Tables 1 and 2.

5.1 Brazilian Portuguese Fake Review Detection dataset

Following the procedures presented in Salminen et al. (2022), we generated a fake reviews dataset in Brazilian Portuguese, based on the Olist dataset presented in Section 4.1. The generated dataset is available online at `https://github.com/cristianomg10/fake-reviews-ptbr-dataset/`.

5.2 Classification results

The results are presented considering the text vectorization methods. In this subsection, we report the results in the following order: Bag-of-Words (BoW), TF-IDF, and BERT. We fixed the number of dimensions in 500 for Bow and TF-IDF, while BERTimbau solely provides 768-dimension representations.

5.2.1 *Bag-of-Words (BoW)*

Figs. 3 to 6 show the confusion matrix obtained using BoW and the respective classifiers. In Fig. 3, we see the results for Logistic Regression. We see that it obtained interesting results regarding true positives, true negatives, and false positives. Only 434 reviews were confused by the classifier.

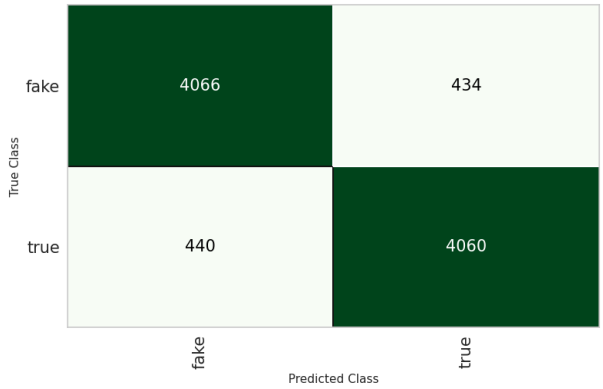

Figure 3: Confusion matrix using Logistic Regression and Bag-of-Words.

Considering the Decision Tree, it obtained inferior results compared to Logistic Regression, reaching almost double the false positives obtained by the Logistic Regression. Fig. 4 shows the confusion matrix for the combination Decision Tree and BoW. Both true positives and true negatives correspond to around 90% of those obtained by Logistic Regression.

Random Forest, as expected due to its robustness, obtained better results than Decision Tree. However, the results obtained are slightly worse than those of the Logistic Regression. Fig. 5 presents the confusion matrix for the combination Random Forest and BoW.

At last, Fig. 6 shows the confusion matrix for SVM and BoW. It obtained interesting results in terms of true positives and true negatives. However, SVM obtained slightly worse results than Logistic Regression regarding false positives while performing better in terms of false negatives.

To conclude this subsection, Logistic Regression with BoW reached the best results in terms of accuracy, F1, and false positives. Precisely, the respective values are 90.49% $\pm$ 0.40, 90.49% $\pm$ 0.40, and 434.

5.2.2 *TF-IDF*

Figs. 7 to 10 show the confusion matrix obtained using TF-IDF and the respective classifiers. Fig. 7 shows the

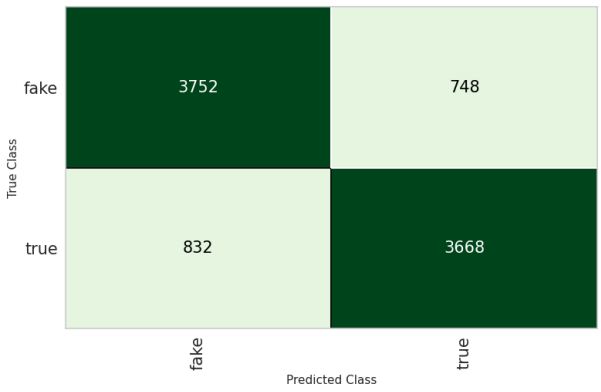

Figure 4: Confusion matrix using Decision Tree and Bag-of-Words.

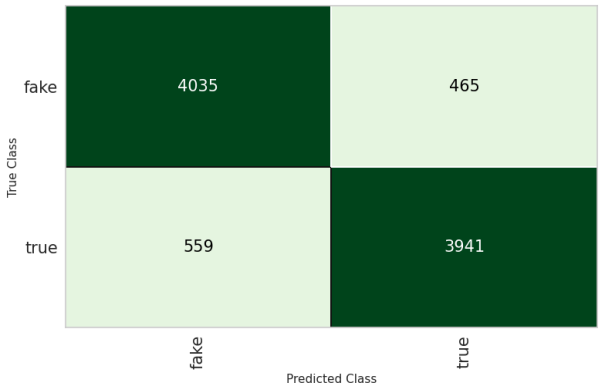

Figure 5: Confusion matrix using Random Forest and Bag-of-Words.

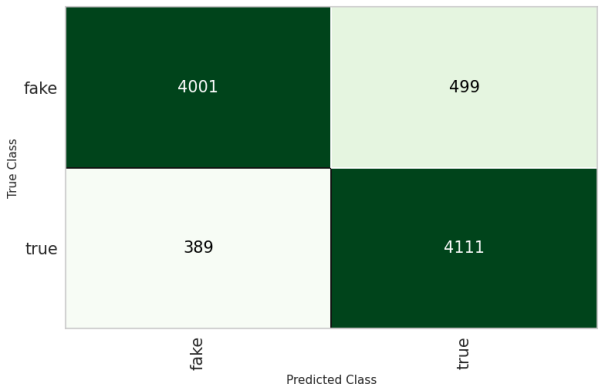

Figure 6: Confusion matrix using SVM and Bag-of-Words.

confusion matrix for Logistic Regression and TF-IDF. We see that Logistic Regression with BoW obtained better results than Logistic Regression and TF-IDF. An increase in false positives and false negatives is noticed in this

scenario.

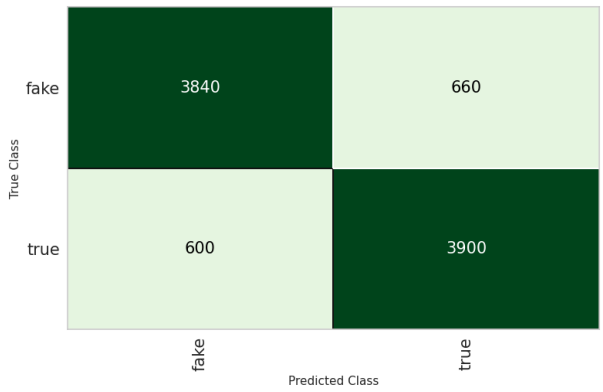

Figure 7: Confusion matrix using Logistic Regression and TF-IDF.

Considering the combination Decision Tree and TF-IDF, Fig. 8 shows the confusion matrix obtained. In opposition to using BoW, the Decision Tree together with TF-IDF obtained better results, decreasing the false negatives and false positives by almost 10%.

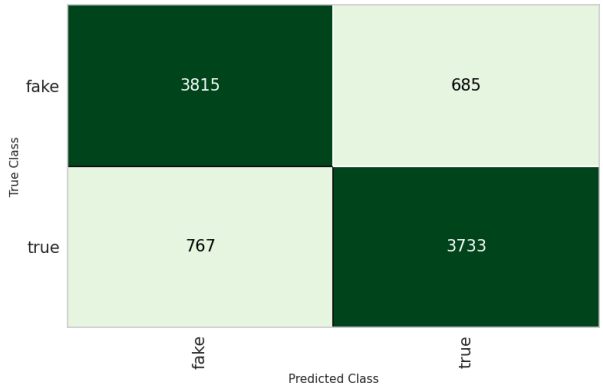

Figure 8: Confusion matrix using Decision Tree and TF-IDF.

Fig. 9 shows the confusion matrix regarding the Random Forest and TF-IDF. Although this combination resulted in interesting results, there was an increase in false positives compared to Random Forest and BoW. In addition, this combination obtained fewer false negatives than the previous approaches, mainly considering the Decision Tree. It is expected since Random Forest is a combination of Decision Trees, being more robust.

Finally, using SVM and TF-IDF, this setting resulted in an increase of 45% in false positives compared to SVM and BoW. Fig. 10 shows the confusion matrix regarding SVM and TF-IDF.

Considering the combination using TF-IDF, Random Forest obtained the best results. The numbers of false

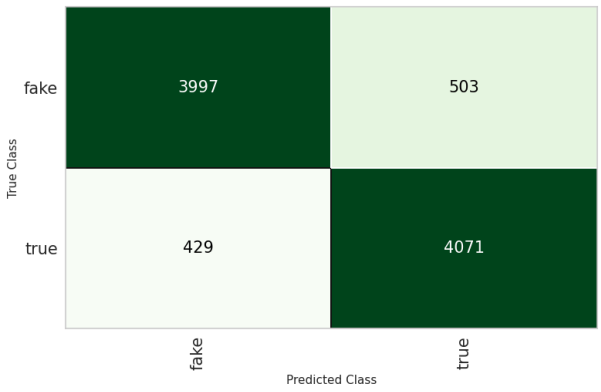

Figure 9: Confusion matrix using Random Forest and TF-IDF.

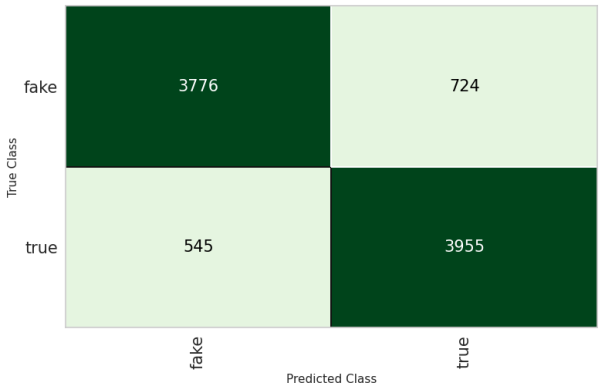

Figure 10: Confusion matrix using SVM and TF-IDF.

positives and false negatives are the smallest for this setting. Random Forest and TF-IDF reached 89.64% of accuracy and 89.56% of F1-Score.

5.2.3 BERT

Figs. 11 to 14 show the confusion matrix obtained using the pre-trained BERT, i.e., BERTimbau, and the respective classifiers. Fig. 11 shows the confusion matrix obtained by using Logistic Regression. We can see an outstanding result considering the low number of false positives and false negatives. Regarding the false positive, it is the lowest value obtained in the experiments.

Regarding the results from the Decision Tree with BERT, shown in Fig. 12, the number of false positives aligns with the previously presented combinations, being considerably high. It reached 589 false positives in the test set, showing that a single Decision Tree is not robust enough for this dataset. In addition, the number of false negatives increased dramatically compared to Logistic Regression and BERT.

Fig. 13 shows the results for Random Forest and BERT. We see that a combination of Decision Trees could reduce to almost a third, compared to the false positives obtained by Decision Tree. However, Random Forest obtained

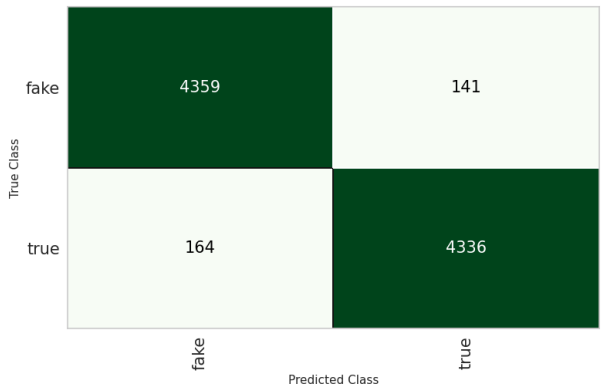

Figure 11: Confusion matrix using Logistic Regression and BERT.

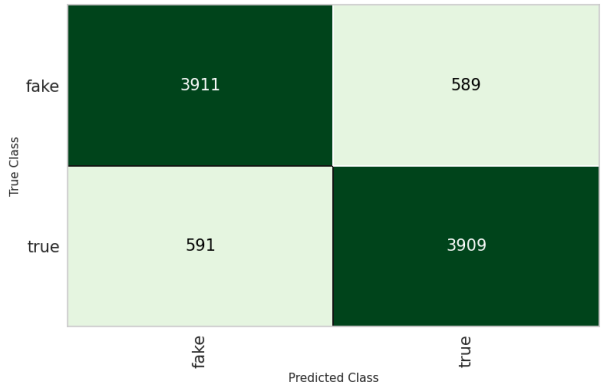

Figure 12: Confusion matrix using Decision Tree and BERT.

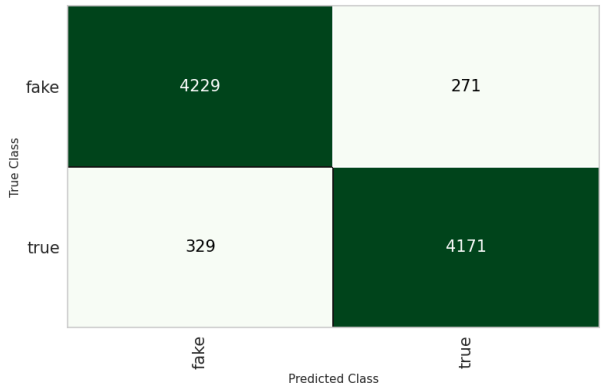

Figure 13: Confusion matrix using Random Forest and BERT.

almost double the false positives obtained by the Logistic Regression.

Finally, Fig. 14 shows the confusion matrix for SVM and BERT. SVM was the closest to Logistic Regression,

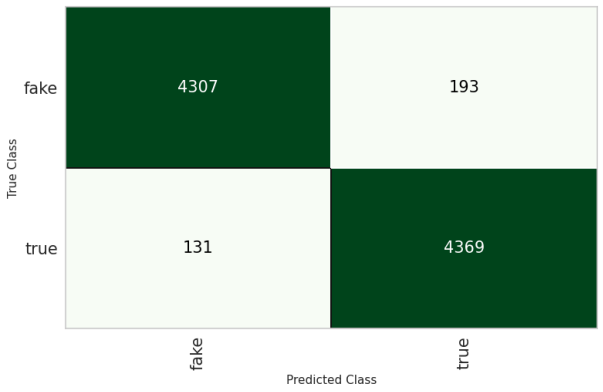

Figure 14: Confusion matrix using SVM and BERT.

obtaining 193 false positives in the test set. In addition, SVM and BERT hit true positives and true negatives more than 10% above the Decision Tree.

Table 1 displays the results, considering a stratified K-fold strategy with 10 folds. The values in bold are the best considering the text vectorization method, while the values with the star (*) are globally the best. We see that, for the proposed dataset, Logistic Regression was the best in two scenarios, i.e., using Bag-of-Words (BoW) and BERT. Interestingly, using TF-IDF, Random Forest was the best classifier, almost 4 percent points on average above the results obtained by Logistic Regression. Furthermore, BERT provided the best representations. It is expected due to its ability to capture contexts instead of only counting words and measuring the importance of words without perceiving their relations.

Table 1: Summarization of the results. Results were obtained using stratified K-fold using 10 folds.

Text Vect.	Classifier	Accuracy (%)	F1-Score (%)
BoW	Logistic Regression	**90.49 ± 0.40**	**90.49 ± 0.40**
BoW	Decision Tree	83.26 ± 0.44	83.24 ± 0.45
BoW	Random Forest	88.90 ± 0.80	88.84 ± 0.83
BoW	SVM	89.87 ± 0.53	89.75 ± 0.63
TF-IDF	Logistic Regression	85.90 ± 0.69	85.88 ± 0.70
TF-IDF	Decision Tree	83.54 ± 0.67	83.52 ± 0.67
TF-IDF	Random Forest	**89.40 ± 0.63**	**89.37 ± 0.64**
TF-IDF	SVM	86.15 ± 0.60	86.08 ± 0.61
BERT	Logistic Regression	**96.55 ± 0.31***	**96.54 ± 0.31***
BERT	Decision Tree	87.02 ± 0.60	87.02 ± 0.60
BERT	Random Forest	93.31 ± 0.43	93.28 ± 0.43
BERT	SVM	96.12 ± 0.49	96.06 ± 0.54

Table 2 shows the results using a stratified holdout strategy. Precisely, the results were obtained from the test set, which corresponded to 30% of the data. Again, Logistic Regression with BERT resulted in the best performance among the evaluated strategies, reaching the smallest number of false positives.

6 Conclusion

Fake reviews have become a concern on e-commerce platforms since they can jeopardize product reputation

Table 2: Results were obtained using the test set in a holdout strategy using 30% of the data for the test.

Text Vect.	Classifier	Accuracy (%)	F1-Score (%)	FP
BoW	Log. Regression	**90.29**	**90.30**	**434**
BoW	Decision Tree	82.44	82.61	748
BoW	Random Forest	88.62	88.74	465
BoW	SVM	90.13	90.01	499
TF-IDF	Log. Regression	86.00	85.91	660
TF-IDF	Decision Tree	83.87	84.01	685
TF-IDF	Random Forest	**89.64**	**89.56**	**503**
TF-IDF	SVM	85.90	85.61	724
BERT	Log. Regression	**96.61***	**96.62***	**141**
BERT	Decision Tree	86.89	86.89	589
BERT	Random Forest	93.33	93.38	271
BERT	SVM	96.40	96.38	193

and lead users to make decisions based on unreal information. Fake review datasets are scarce in the literature. Therefore, this paper presented a fake review dataset in Brazilian Portuguese based on the Olist dataset.

Following the procedure presented in Salminen et al. (2022), we fine-tuned a GPT-2 model to generate the fake reviews. In addition, we evaluated three text vectorization methods, i.e., BoW, TF-IDF, and BERT (BERTimbau), together with four classifiers, i.e., Logistic Regression, Decision Tree, Random Forest, and SVM. Logistic Regression with BERT reached the best values regarding F1-Score, accuracy, and false positives.

Although this paper contributes with resources for the literature, i.e., a GPT-2 model for review generation in Brazilian Portuguese and a fake review dataset in Brazilian Portuguese, it has some limitations. For example, computer-generated reviews do not necessarily have the characteristics of fake reviews. On the other hand, companies such as Yelp leverage an unknown filter algorithm to detect fake reviews (Mohawesh et al., 2021a). This highlights the difficulty of performing research in this area.

In future works, we intend to develop methods for incremental detection of fake reviews, using the reviews as a text stream (Garcia et al., 2025a), which is a more realistic scenario for online e-commerce platforms (Gama et al., 2014). In a text stream, texts arrive individually, and an incremental classifier must learn from and discard the new input text. Furthermore, in such scenarios, it is possible to evaluate potential concept drifts in the fake reviews.

References

Akesson, J. et al. (2023). The Impact of Fake Reviews on Demand and Welfare, *Technical report*, National Bureau of Economic Research. http://dx.doi.org/10.3386/w31836.

Berrou, B. K., Al Kalbani, K., Antonijevic, M., Zivkovic, M., Bacanin, N. and Nikolic, B. (2023). Training a Logistic Regression Machine Learning Model for Spam Email Detection using the Teaching-Learning-based-Optimization Algorithm, *Proceedings of the 1st International Conference on Innovation in Information Technology and Business (ICIITB 2022)*, Vol. 104, p. 306. http://dx.doi.org/10.2991/978-94-6463-110-4_22.

BrightLocal (n.d.). Why are fake reviews a problem? Available at: https://www.brightlocal.com/learn/review-management/fake-reviews/why-are-fake-reviews-a-problem/.

Cao, C. (2023). The Impact of Fake Reviews of Online Goods on Consumers, *BCP Business & Management* . http://dx.doi.org/10.54691/bcpbm.v39i.4208.

Cox, D. R. (1958). The Regression Analysis of Binary Sequences, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **20**(2): 215–232. Available at https://www.jstor.org/stable/2983890.

da Costa, P. T. G. C., de Almeida, J. F. F. A., Fernandes, J. M. and Ortega, L. M. (2021). E-commerce no Brasil: Revisão Sistemática de Literatura de 2011 a 2021, *Brazilian Journal of Business* **3**(4): 2969–2982. http://dx.doi.org/10.34140/bjbv3n4-014.

da Cruz Machado Benatti, N. C. (2017). Aprendizagem de Máquina Aplicada a Investimentos Imobiliários, *Simpósio de Métodos Numéricos em Engenharia - UFPR* . Available at https://hdl.handle.net/1884/93040.

de Almeida Teodoro, L. and Kappel, M. A. A. (2020). Aplicação de Técnicas de Aprendizado de Máquina para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil, *Revista Brasileira de Informática na Educação* . http://dx.doi.org/10.5753/rbie.2020.28.0.838.

de Andrade, M. C. F. and Silva, N. G. (2017). O comércio eletrônico (e-commerce): um estudo com consumidores, *Perspectivas em Gestão & Conhecimento* **7**(1): 98–111. http://dx.doi.org/10.21714/2236-417X2017v7n1.

de Melo Cruz, W. L. (2021). Crescimento do E-commerce no Brasil: Desenvolvimento, Serviços Logísticos e o Impulso da Pandemia de Covid-19, *GeoTextos* . http://dx.doi.org/10.9771/geo.v17i1.44572.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* . http://dx.doi.org/10.18653/v1/N19-1423.

Ebit (2021). 44ª ed. webshoppers. Available at https://eyagencia.com.br/wp-content/uploads/2021/09/Webshoppers_44-relatorio-2021-resultados-ecommerce-ebit.pdf.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014). A Survey on Concept Drift Adaptation, *ACM Computing Surveys (CSUR)* **46**(4): 1–37. http://dx.doi.org/10.1145/252381.

Garcia, C. M. et al. (2025a). Concept Drift Adaptation in Text Stream Mining Settings: A Systematic Review, *ACM Transactions on Intelligent Systems and Technology* . http://dx.doi.org/10.1145/370492.

Garcia, C. M. et al. (2025b). Improving Sampling Methods for Fine-tuning SentenceBERT in Text Streams, *Lecture Notes in Computer Science (LNCS, volume 15319)* . http://dx.doi.org/10.1007/978-3-031-78495-8_28.

Godoy, L. C., Farkouh, M. E., Austin, P. C., Shah, B. R., Qiu, F., Sud, M., Wijeysundera, H. C., Mancini, G. J. and Ko, D. T. (2023). Predicting left main stenosis in stable ischemic heart disease using logistic regression and boosted trees, *American Heart Journal* **256**: 117–127. http://dx.doi.org/10.1016/j.ahj.2022.11.004.

Harris, Z. S. (1954). Distributional Structure, *Word* **10**(2-3): 146–162. http://dx.doi.org/10.1080/00437956.1954.11659520.

He, S., Hollenbeck, B. and Proserpio, D. (2022). The Market for Fake Reviews, *Marketing Science* **41**(5): 896–921. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3664992.

Leite, D. R. A. et al. (2023). Método de aprendizagem de máquina para classificação da intensidade do desvio vocal utilizando random forest, *Journal of Health Informatics* . Available at https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/814.

Mohawesh, R. et al. (2021a). Analysis of Concept Drift in Fake Reviews Detection, *Expert Systems with Applications* **169**: 114318. http://dx.doi.org/10.1016/j.eswa.2020.114318.

Mohawesh, R. et al. (2021b). Fake Reviews Detection: A Survey, *IEEE Access* **9**: 65771–65802. http://dx.doi.org/10.1109/ACCESS.2021.3075573.

Mota, M. d. O. (2021). Estudo de Caso sobre Segurança em E-commerce, *Pontifícia Universidade Católica de Goiás* . Available at https://repositorio.pucgoias.edu.br/jspui/handle/123456789/3206.

Nascimento, R. (2011). E-Commerce no Brasil: Perfil do Mercado e do E-consumidor Brasileiro. Available at: https://bibliotecadigital.fgv.br/dspace/handle/10438/8182.

Nick, T. G. and Campbell, K. M. (2007). Logistic Regression, *Topics in Biostatistics* pp. 273–301. http://dx.doi.org/10.1891/9781617050992.0039.

Novaes, A. (2016). *Logística e Gerenciamento da Cadeia de Distribuição*, Elsevier Brasil.

Olist and Sionek, A. (2018). Brazilian E-Commerce Public Dataset by Olist. http://dx.doi.org/10.34740/KAGGLE/DSV/195341.

Otieno, D. O., Namin, A. S. and Jones, K. S. (2023). The Application of the BERT Transformer Model for Phishing Email Classification, *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, IEEE, pp. 1303–1310. http://dx.doi.org/10.1109/COMPSAC57700.2023.00198.

Ott, M., Choi, Y., Cardie, C. and Hancock, J. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination, *Journal of Retailing and Consumer Services* . http://dx.doi.org/10.48550/arXiv.1107.4557.

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* **12**: 2825–2830. Available at https://dl.acm.org/doi/pdf/10.5555/1953048.2078195.

Prusty, S., Patnaik, S. and Dash, S. K. (2022). SKCV: Stratified K-fold Cross-validation on ML Classifiers for Predicting Cervical Cancer, *Frontiers in Nanotechnology* **4**: 972421. http://dx.doi.org/10.3389/fnano.2022.972421.

Radford, A. et al. (2019). Language Models are Unsupervised Multitask Learners, *OpenAI blog* **1**(8): 9. Available at https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Runchi, Z., Liguo, X. and Qin, W. (2023). An ensemble credit scoring model based on logistic regression with heterogeneous balancing and weighting effects, *Expert Systems with Applications* **212**: 118732. http://dx.doi.org/10.2139/ssrn.4167821.

Salminen, J., Kandpal, C., Kamel, A., Jung, S. and Jansen, B. (2022). Creating and Detecting Fake Reviews of Online Products, *Journal of Retailing and Consumer Services* . http://dx.doi.org/10.1016/j.jretconser.2021.102771.

Salton, G. and Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval, *Information Processing & Management* **24**(5): 513–523. http://dx.doi.org/10.1016/0306-4573(88)90021-0.

Solaiman, I. et al. (2019). Release Strategies and the Social Impacts of Language Models, *arXiv preprint arXiv:1908.09203* . http://dx.doi.org/10.48550/arXiv.1908.09203.

Souza, F., Nogueira, R. and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese, *9th Brazilian Conference on Intelligent Systems, BRACIS*. http://dx.doi.org/10.1007/978-3-030-61377-8_28.

Thuma, B. S. et al. (2023). Benchmarking Feature Extraction Techniques for Textual Data Stream Classification, *2023 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 1–8. http://dx.doi.org/10.1109/IJCNN54540.2023.10191369.

Topal, M. O., Bas, A. and van Heerden, I. (2021). Exploring Transformers in Natural Language Generation: GPT, BERT, and XLNet, *ArXiv 2102.08036* . https://doi.org/10.48550/arXiv.2102.08036.

Tufail, H. et al. (2022). The Effect of Fake Reviews on e-Commerce During and After Covid-19 Pandemic: SKL-Based Fake Reviews Detection, *Inst. of Electrical and Electronics Engineers* . http://dx.doi.org/10.1109/ACCESS.2022.3152806.

Valant, J. (2015). Online Consumer Reviews: The Case of Misleading or Fake Reviews, *European Parliamentary Research Service* . Available at https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2015)571301.

Which? (2020). The Real Impact of Fake Reviews: a Behavioural Experiment on how Fake Reviews Influence Consumer Choices, *The Behaviouralist* . Available at: https://media.product.which.co.uk/prod/files/file/gm-32c0deb8-a98c-4a3f-85db-8393284106a6-fake-reviews-summary-report.pdf.