

Authentic and Fake Reviews Recognition on E-Commerce Websites through Sentiment Analysis and Machine Learning Techniques

Kian Nimgaz Naghsh^{a*}, Ali Asghar Pour Haji Kazem^b

^a Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran, kian_naghsh@yahoo.com

^b Department of Software Engineering, Faculty of Engineering and Natural Science, Istinye University, Istanbul, Turkey, ali.pourhaji@istinye.edu.tr

ABSTRACT


The proliferation of e-commerce has led to an overwhelming volume of customer reviews, posing challenges for consumers who seek reliable product evaluations and for businesses concerned with the integrity of their online reputation. This study addresses the critical problem of detecting fake reviews by developing a comprehensive framework that integrates Natural Language Processing (NLP) and machine learning techniques. Our methodology centers on sentiment analysis to discern the emotional valence of reviews, coupled with Part-of-Speech (PoS) tagging to analyze linguistic patterns that may signal deception. We meticulously extract a rich set of textual and statistical features, providing a robust basis for our predictive models. To enhance classification performance, we strategically employ both traditional machine learning algorithms and powerful ensemble techniques. Experimental results underscore the efficacy of our approach in detecting fraudulent reviews. We achieved a notable F1-Score of 82.9% and an accuracy of 82.6%, demonstrating the potential to safeguard consumers from misleading information and protect businesses from unfair practices.

Keywords— Fake Review, Authentic Review, E-Commerce websites, Sentiment Analysis, Machine Learning

1. Introduction

With the rapid development of the Internet, the impact of online reviews increases continuously. E-commerce Portals are getting increasingly popular to share customer views [1]. These developments have severely altered how opinions and reviews are shared on e-commerce websites and cyberspace. Online reviews are comments, tweets, posts, and opinions on different online platforms like review sites, news sites, e-commerce sites, or any other social networking sites [2]. In other words, reviews are an individual's thoughts or experiences about a product or service after online shopping. Similarly, customers got used to going through reviews available before purchasing a product. In buying a product from online shops, a person must read all the reviews written by others to check if the product

is the case and suitable or not. Thus, reviews of products have become an essential source of information for buyers. However, online reviews also have a negative effect. Because of this tendency of customers, online reviews have become a target for spammers. Due to the high profit of e-commerce websites, they are a base for fraudulent activities like other business platforms. In some cases, spammers who write fake reviews about a product or service do this to promote a product or demote another to damage its reputation. Fake reviews are also known as deceptive opinions, spam opinions, phony reviews, or spam reviews [3]. Consequently, they can cause financial loss for merchandisers and service providers because negative fake reviews can damage their brand reputation. They also cause companies to make more profit by posting fake positive reviews [4]. Today, millions of e-commerce websites are working around the world. We can see

 <http://dx.doi.org/10.22133/ijwr.2024.425201.1194>

Citation K. Nimgaz Naghsh, A. Asghar Pour Haji Kazem, " Authentic and Fake Reviews Recognition on E-Commerce Websites through Sentiment Analysis and Machine Learning Techniques ", *International Journal of Web Research*, vol.6, no.2, pp.119-131, 2023, doi: <http://dx.doi.org/10.22133/ijwr.2024.425201.1194>.

*Corresponding Author

Article History: Received: 14 September 2023; Revised: 23 December 2023; Accepted: 28 December 2023.

Copyright © 2022 University of Science and Culture. Published by University of Science and Culture. This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 International license (<https://creativecommons.org/licenses/by-nc/4.0/>). Noncommercial uses of the work are permitted, provided the original work is properly cited.

tons of reviews written on the products. It is estimated that approximately 8–15% of these platforms' reviews are fake [5]. However, how can we be sure about their authenticity? These fake reviews are highly complex to be understood by machines [5]. Regarding this fact, because of the importance of the subject, numerous researchers have been working on finding new methods to understand if the information written as reviews is authentic or fake. It is also stated in several research projects that competitors hire spammers to write fake reviews. They have found the trustworthiness of online reviews to be questionable. For instance, Yelp, an online service provider, estimated that 20% of the reviews on their website are faked by paid writers [6].

It is also challenging to know if a customer is real or has used the product he or she is reviewing and is giving genuine feedback about that product or service. Because just a few sites restrict users to post reviews only for purchased items. Thus, not all reviews on the Internet express authentic post-purchase experiences [7]. Some people are paid to comment on products on websites as content producers. Some others could be fake and written based on imagination. Users could post fake reviews to gain status in the community or simply for fun [8]. Alternatively, they can be Computer-Generated reviews [3] by robots which is more possible nowadays. Due to advancements in technology and Natural Language Processing (NLP) and Artificial Intelligence (AI), text-producing machines are more accessible these days, and crowd workers rarely are employed to write fake reviews on websites.

Some researchers in this field call this phenomenon a Reputation System. The reputation systems aim at helping consumers in deciding whether to negotiate with a given party [9]. Most of the available reputation models rely on numeric data, such as ratings on e-commerce websites. Also, most of the reputation models focused only on the overall ratings of products without considering reviews provided by buyers. Fake reviews, either written by spammers or generated by machines (AI), are authentic-appearing, deceptive, and hard to differentiate. So, customers reviewing these comments could be misled in their decision to buy the right product or service. This may negatively affect online shopping and primarily small online businesses. We can differentiate fake reviews based on their detail level and writing styles. Thus, Natural Language Processing methods such as Sentiment Analysis (SA) and Supervised Machine Learning techniques will positively affect fake review detection.

Supervised Machine Learning is a subcategory of Machine Learning that uses a dataset to train the algorithm to predict the outcomes accurately or

classify the data. This approach uses a target feature as a label that must be predicted. Furthermore, the mentioned dataset should be split into train and test datasets to test the model's performance at the final stage. This paper examines whether authentic and fake reviews are distinguishable using Supervised Machine Learning based on Natural Language Processing techniques and Sentiment Analysis. Although many researchers have made valuable contributions to this field, there are some gaps in the fake reviews detection. For instance, most studies focus on detecting fake reviews using just Yelp datasets ([10], [11]). Even though it is a gold-standard dataset, many other datasets and areas need to be investigated too. Moreover, there is no standard labeled dataset specialized in e-commerce reviews, even if it's considered a trending area in the last decades.

In this document, to solve this malignant problem, we explore different research studies about detecting fake reviews. In this work, we have used Sentiment Analysis, study and analysis of people's opinions and emotions about products or services and giving it a polarity and subjectivity score. Moreover, we have applied a very efficient PoS-tagging-based aspect extraction methods for a better extraction of aspects and polarity from reviews, various Ensemble Learning algorithms, which hardly are used in similar works which use traditional Machine Learning algorithms. Furthermore, another distinguishing point of this work that rarely can be seen in previous studies is that multiple executions are applied for proposed algorithms, which can lead the results to a better point. Also, we prove that default parameters of the proposed Machine Learning algorithms can be improved by the numerous execution of the algorithm.

We have organized the rest of this paper as follows: in Section 2, we introduce some related works in this field; in Section 3, we explain our proposed approach for gathering data and Sentiment Analysis; in Section 4, we present our Machine Learning model; in Section 5, we will evaluate our experiment and will see experimental results; and finally, discussion, conclusions, and directions for future work are given in Section 6 and 7 respectively.

2. Related Works

Much work has been done on fake review detection, spam filtering, and fake news detection [12]. Different approaches, such as Supervised, Unsupervised, and Semi-Supervised learning algorithms, have been proposed, and different data mining and analysis techniques have been used. Also, researchers have used various datasets to find solutions and models to differentiate fake reviews

from genuine ones. One of the preliminary and leading works on this subject is done by Ott et al. [10]. They claimed that detecting deceptive opinion spam is well beyond the capabilities of human judges, most of whom perform roughly at-chance. They developed the first large-scale dataset. They used that dataset containing 400 valid and 400 fake reviews. They found that although standard n-gram-based text categorization is the best individual detection approach, a combination approach using psycholinguistically-motivated features and n-gram features can perform slightly better. The study conducted by [13] used the "fsQCA" or "fuzzy set Qualitative Comparative Analysis" method on reviews from the Yelp dataset to identify fake and true reviews. They have identified two types of review patterns, authentic and fake, based on configurations among reviewers and review content elements. They have also explained what fake review is with IMT as theoretical background. Based on the fsQCA method, the results have identified the combinations of configurations for authentic and fake reviews. In [3], authors have used two language models, ULMFiT and GPT-2, to generate fake product reviews based on an Amazon e-commerce dataset. They used Amazon Original reviews plus their computer-generated reviews to analyze data to detect fake reviews based on some models such as OpenAI and RoBERTa. Banerjee et al. proposed a classification model for authentic and fake reviews using supervised learning [7]. They suggested that four linguistic clues could help to distinguish between authentic and fake reviews. They created a dataset consisting of 900 genuine reviews and 900 fake reviews for some popular hotels in Asia. In [9] and [14], authors proposed similar methods to analyze a dataset of movie reviews. They presented supervised learning algorithms and sentiment classification using and not using stop-words. They studied the accuracy of sentiment classification algorithms such as KNN, K*, NB, SVM, and DT-J48. Some other researchers in the past have proposed unlabeled approaches using Positive-Unlabeled Learning (PU Learning) [15]. Their Exploratory results show that PU Learning not only significantly outperforms supervised learning but also detects many potentially fake reviews hidden in the unlabeled set. On the other hand, some authors propose a graph-based method such as [16]. The main idea of this model is to depict relationships among entities and analyze the importance of features by calculating weights based on feature fusion techniques. Consequently, they determined the most practical combination of weighted features. After that, the feature selection is applied using IG (Information Gain) and TF-IDF, and the most compelling features are selected by applying a well-known classifier (SVM, NB, and DT). Authors in [17] have presented an approach for detecting spam and not-useful

reviews. They also used KNN, SVM, and NB classifiers to reviews classification and prioritize them based on their weight (confidence). In [5], authors have used the Yelp dataset for fake review detection using Convolutional Neural Network (CNN) as a Deep Learning technique and Long Short Term Memory (LSTM). They used aspects instead of complete detailed review text for their analysis. Sentiments of aspects were calculated using PoS-tagging and Senti-WordNet. Thus such extracted characteristics are fed into CNN and LSTM models for aspect replication and fake review detection. They proposed an approach with high accuracy of more than 90%. In [1], authors used the n-gram feature, a contiguous sequence of n items from a given text, for their approach. They analyzed and trained their model with both traditional and ensemble machine learning algorithms and concluded that ensemble methods work better. They have utilized the Gold-Standard dataset obtained by [10] for their experiments. In [18], a dataset developed that contained Urdu and Roman Urdu reviews. Their results showed that text categorization with an SVM classifier is the most suitable approach for fake reviews detection using the n-gram approach. And, in [6], researchers have proposed readability tests (difficulty level of a text to be read and understood) on review text as features for fake review detection. They examined various tests used to measure a text's readability.

3. Proposed Approach

We consider fake review detection quite a challenging issue. From the human perspective, it is too hard for humans to manually distinguish between real and phony reviews without any tools for text analysis. So, diverse methods have been devised for detecting spam reviews. Researchers ([1] and [4]) have adopted three approaches:

- The first approach focuses on the content of reviews written by a reviewer and attempts to realize their affinity. Some of the features used in this method are the length of the comments, the number of words in each part of speech (nouns, verbs, adjectives, etc.), sentiment polarity, and other linguistic characteristics.
- The second strategy is based on the reviewer. It concentrates on the behavior of reviewers and considers information about users such as reviewer id, timing behavioral statistics about reviewers, and all reviews noted by them. Then the reviewer is classified as a spammer or non-spammer based on the mentioned features.
- Last but not least, the third strategy is the product-centric technique. This method

explicitly underlines the details related to each product. So we use some product features to develop our models, such as the mean rating given to the product, the product's price, and the product's sales rank.

In this paper, we use the first approach (content-based) plus a robust dataset to analyze our data with Natural Language Processing techniques such as Sentiment Analysis. Then we train our model with Supervised Machine Learning algorithms to classify our data as fake or authentic. Then we compare performance to detect computer-generated fake reviews.

3.1. Dataset

Researchers have used different datasets for review Sentiment Analysis and fake review detection. Even some of them have produced their own-generated datasets. However, not all of these datasets are trustworthy and have their tribulations. In this paper, we analyze a dataset from [3] to accomplish our goal. The dataset includes about 40,000 reviews in total, a good number of samples for a text classification task. It contains 20,000 artificially-generated (fake) reviews. It also has 20,000 real reviews authored by humans, original samples from the Amazon dataset from the top Amazon categories with the most product reviews [19]. In Figure 1 we see a histogram that represents the data about the number of words in each text and the count of texts with that amount of words.

3.2. Sentiment Analysis

As we discussed before, customers' reviews and opinions significantly affect individual decision makings in their activities, such as online shopping. Sentiment Analysis has recently become one of the most exciting text analysis subjects due to its advantageous commercial benefits. Sentiment Analysis, also called Opinion Mining [20], is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. One of the principal issues encountering Sentiment Analysis is extracting opinions and emotions inside the text and determining polarity and subjectivity scores for that. In the next phase, it desires to classify opinions based on those scores. In this paper, we use Sentiment Analysis to classify the text as a positive, negative, or neutral text by giving it a score and using it as one of the features to train our model. In Figure 2, we can see the histogram representing the distribution of polarity score and the count of texts with that polarity. Moreover, efficient PoS-tagging-based aspect extraction techniques are applied to extract aspects and polarity from reviews. In the next session, we will discuss more PoS-tagging.

3.3. PoS (Part-of-Speech) tagging

We define PoS-tagging as the process of a word demarcation in a corpus (text). In other words, we determine the word as a particular part of speech based on its definition and context. In general, there are nine primary parts of speech in English: noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, and interjection. However, there can be other sub-groups in which a word can stand. As proposed earlier, this paper endeavors to differentiate authentic and phony reviews based on Sentiment Analysis and Machine Learning algorithms. In the Machine Learning process, we need features for our learning process. Here, we use the PoS tags mentioned above, as shown in Table 1, to categorize all words in a review text to make new features in our dataset [21].

4. Machine Learning model

Figure 1 displays the architectural diagram of our project. In the first phase, as we discussed earlier, we gather the data and prepare the dataset. Then, the pre-processing phase is done. Data pre-processing plays a very significant role in supervised learning models. So, after Sentiment Analysis and PoS-tagging, for one example, we use MinMaxScaler provided by the sklearn library to transform features by scaling each feature to a given range. This function scales each feature separately to be in the given range on the training set (for example, between zero and one). After the pre-processing phase, we select features to train our model. There is no standard number for the number of features in the classification problem. So it depends on the problem and correlations between data and our target feature. Here we eliminate some features not only for their low correlation but also to diminish the processing load of our algorithm. After a comprehensive examination of the features, our final features are as follows in Table 2.

Then the data is divided using the sklearn train-test-split library to split the dataset in 80:20 for training and testing the model. After that, we train our model using labeled train dataset. We use various supervised machine learning algorithms as follows: Gaussian Naïve Bayes (GNB), K-Nearest Neighbors Classifier (KNN), Decision Tree Classifier (DTC), Extra Tree Classifier (ETC), Stochastic Gradient Descent Classifier (SGDC), Ada-Boost Classifier (ABC), Logistic Regression (LR), Support Vector Classifier (SVC), Random Forest Classifier (RF), Gradient Boosting Classifier (GB), eXtreme Gradient Boosting Classifier (XGBoost), and Histogram-based Gradient Boosting Classifier (HBGB). Our target feature is the review type. It is a two-class feature: CG (Computer-Generated) or OR (Original Review). However, we have changed it to a binary class, either 0 or 1. We perform Hyper-parameter

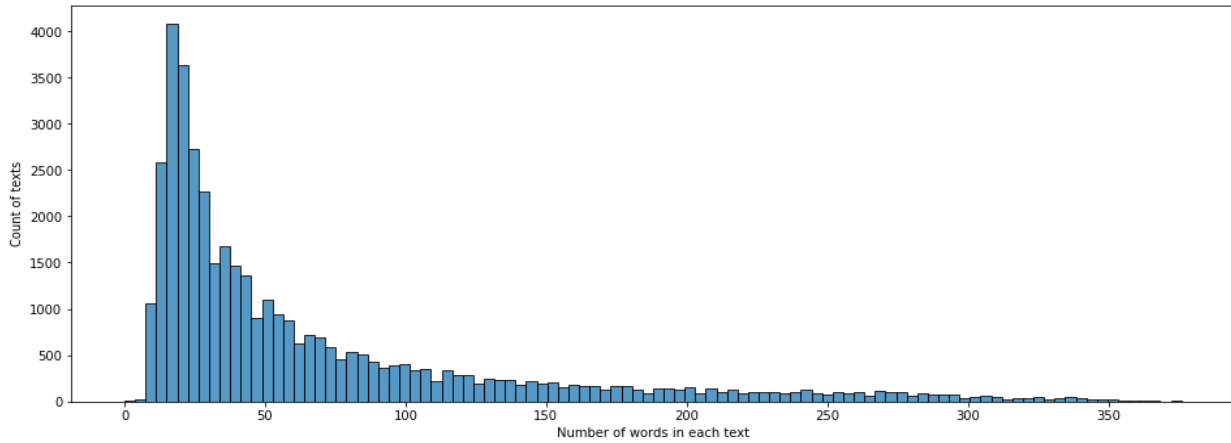


Figure. 1. Histogram that represents the data about the number of words in each text

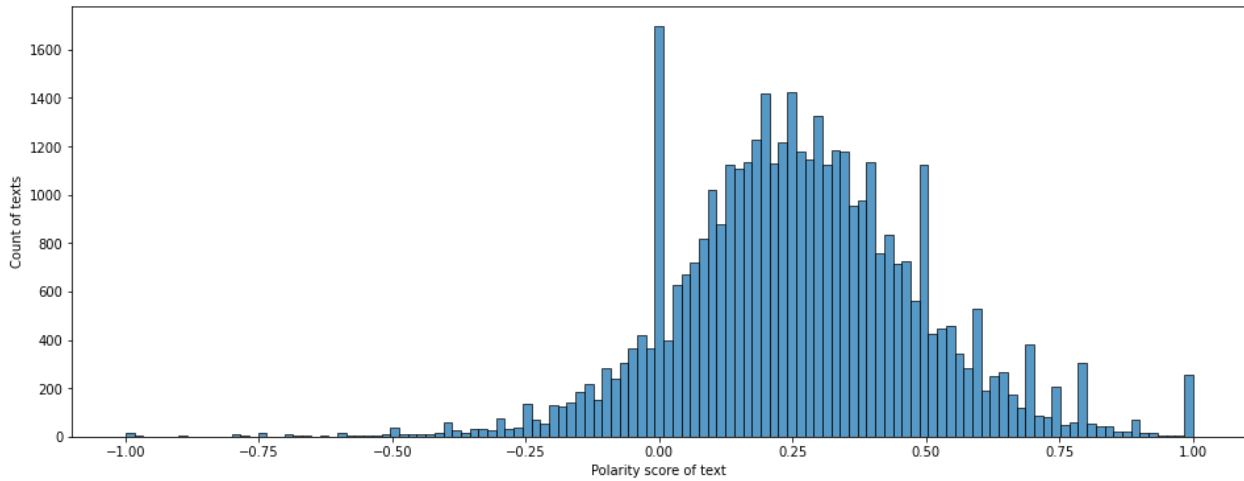


Figure. 2. Histogram that represents the distribution of polarity score

Table 1. Synonym and meaning of PoS-tags

CC: Coordinating conjunction	PRP\$: Possessive pronoun
CD: Cardinal number	RB: Adverb
DT: Determiner	RBR: Adverb, comparative
EX: Existential there	RBS: Adverb, superlative
FW: Foreign word	RP: Particle
IN: Preposition or subordinating conjunction	SYM: Symbol
JJ: Adjective	TO: to
JJR: Adjective, comparative	UH: Interjection
JJS: Adjective, superlative	VB: Verb, base form
LS: List item marker	VBD: Verb, past tense
MD: Modal	VBG: Verb, gerund or present participle
NN: Noun, singular or mass	VBN: Verb, past participle
NNS: Noun, plural	VBP: Verb, non-3rd person singular present
NNP: Proper noun, singular	VBZ: Verb, 3rd person singular present
NNPS: Proper noun, plural	WDT: Wh-determiner
PDT: Pre-determiner	WP: Wh-pronoun
POS: Possessive ending	WP\$: Possessive wh-pronoun
PRP: Personal pronoun	WRB: Wh-adverb

tuning to get optimum results. We executed the model with twelve different Machine Learning algorithms, as shown in Table 3. We used a mixture of traditional Machine Learning algorithms such as tree classifiers and Linear Classifiers and Ensemble and Boosting algorithms. Ensemble and boosting methods are novel methods that are currently being used in Machine Learning research. They use multiple learning algorithms and integrate those base estimators, and in most cases, get better performance than traditional learning algorithms. We can increase the number of epochs with different parameters to find optimal amounts for Hyper-parameters and, consequently, better results and accuracy; we call this process Parameter Tuning or Hyper-parameter Optimization. Then we check each classifier's predicted data's accuracy on test data. Finally, we use a confusion matrix to evaluate the performance of each classifier. Our evaluation criterion is F1-Score and Accuracy Score which calculates the metric of the algorithm in percentage.

5. Evaluation and Experimental Results

After training our labeled data, the next step is to

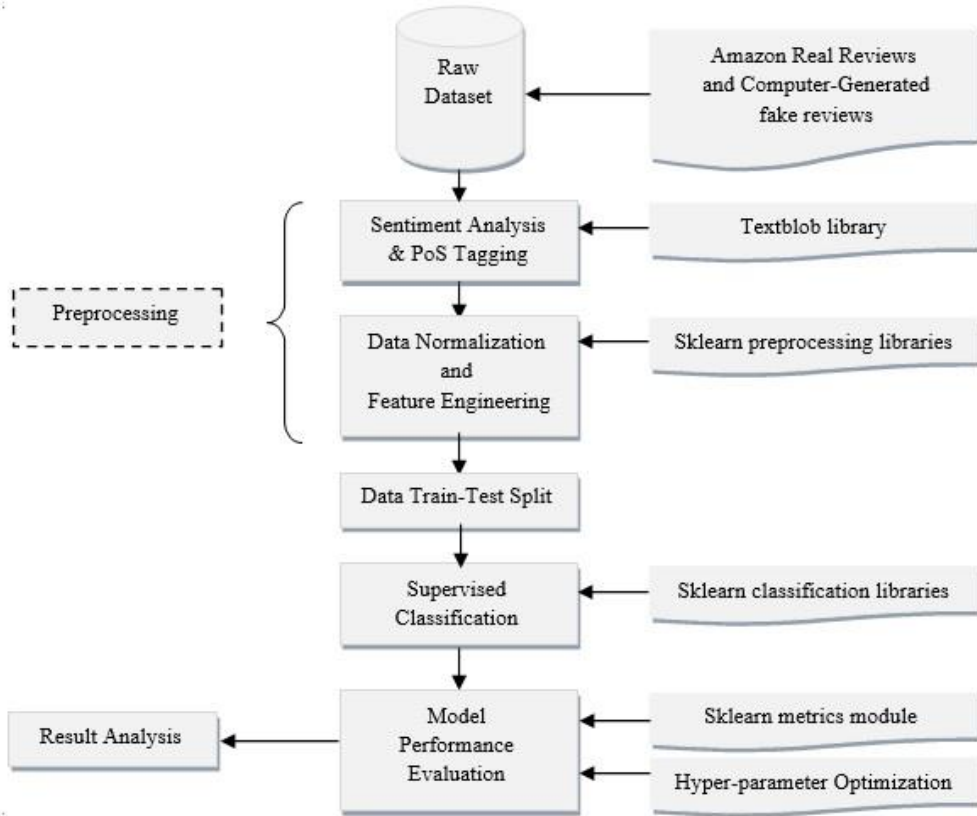


Figure 3. Architectural diagram of our Machine Learning model

predict the model's output using the test dataset. Afterward, we will organize the results in a two-dimensional matrix called the confusion matrix as Table 4. We define each element of the confusion matrix as follows:

TP: True Positive; a review that has been predicted as fake and was labeled as fake.

TN: True Negative; a review that has been predicted as real and was labeled as real.

FP: False Positive; a review that has been predicted as fake and was labeled as real.

FN: False Negative; a review that has been predicted as real and was labeled as fake.

Now, based on these definitions and confusion matrix, we can present all five performance evaluation metric formulas as follows (Equ(1) to (5)):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (4)$$

Table 2. All features selected to train the proposed model

Feature	Description
Number_of_words	Number of words of a review
Length	Length of the review
Rating	Mean rating of the review
Polarity	Polarity of the review in range [-1,1]
Subjectivity	Subjectivity score in range [0,1]
Adjectives	Number of Adjectives
Nouns	Number of Nouns
Verbs	Number of Verbs
Determiners	Number of Determiners
Adverbs	Number of Adverbs
Modals	Number of Modals
Conjunctions	Number of Conjunctions
Pronouns	Number of Pronouns

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

We have used Python 3.8 and Jupyter Notebook platform as an IDE (Integrated Development Environment) to execute our project because of their ease of use and the broad range of online support available. For evaluating the performance of each of these models, we used a confusion matrix and the main classification evaluation metric, F1-Score, for evaluating the performance of the classifiers. Because as we can see in the equations above, the F1-Score considers all parameters of the confusion

Table 3. Results of experiments in our work in comparison with similar works

<i>Algorithm (sklearn Library)</i>	<i>Primary parameters' best value in comparison with their default values</i>	<i>Maximum Accuracy-Score</i>	<i>Max F1-Score in Our work</i>	<i>Max F1-Score in Others works</i>
GNB	Default: var_smoothing=1e-09 Best result: var_smoothing=10	53.9%	65.6%	[7]: 56.4% [22]: 63.1%
KNN	Default: n_neighbors=5 Best result: n_neighbors=43	73.3%	74.8%	[23]: 69.2%
ETC	Default: max_features=sqrt Default: max_depth=None Best result: max_features=13 Best result: max_depth=13	75.6%	77.3%	-
DTC	Default: max_depth=None Best result: depth=11	77.1%	78.0%	[24]: 70.3% [25]: 76.9%
ABC	Default: n_estimators=50 Default: learning_rate=1.0 Best result: estimators=65 Best result: learning=1.3	77.6%	78.1%	[6]: 78.0% [23]: 74.9%
LR	Default: C = 1.0 Best result: C = 69.0	78.8%	79.1%	[7]: 70.9% [6]: 68.0% [24]: 70.3%
SVC	Default: kernel=rbf Best result: kernel=rbf	78.7%	79.5%	[6]: 68.0% [7]: 67.8% [24]: 76.1% [26]: 79.1% [25]: 69.9%
SGD	Default: alpha=0.0001 Best result: alpha=0.00001	78.8%	79.6%	-
GB	Default: learning_rate=0.1 Default: n_estimators=100 Best result: learning_rate=0.6 Best result: n_estimator=60	73.3%	81.7%	[6]: 78.0%
RF	Default: n_estimators=100 Default: max_depth=None Best result: n_estimators=180 Best result: max_depth=33	81.6%	82.0%	[7]: 67.7% [23]: 77.0% [24]: 73.6% [25]: 67.3%
XGBC	Default: learning_rate=0.3 Best result: learning_rate=0.1	82.3%	82.4%	-
HBGB	Default: learning_rate=0.1 Default: max_iter=100 Best result: learning_rate=0.2 Best result: max_iter=500	82.6%	82.9%	-

Table 4. Confusion Matrix

	<i>Predict Fake</i>	<i>Predict Real</i>
<i>Actually Fake</i>	TP	FN
<i>Actually Real</i>	FP	TN

matrix (TP, TN, FP, and FN) and all other metrics (Accuracy, Precision, Recall, and Specificity). So, we evaluate and compare our model with the F1-Score. However, the Accuracy Score for all the algorithms, calculated for better comparison.

Table 3 shows the results. It can be observed that the Hist-Based GB Classifier has explicitly better accuracy than other models, with a score of 82.9%. The Python

code run in Jupyter Notebook for HBGB Classifier is shown in Figure 4.

Moreover, in terms of execution time, the Hist-Based GB Classifier was the fastest algorithm in comparison with all others. It is also evident that our proposed techniques outperform similar researches. Moreover, we can conclude from the results that the default values of the main parameters of each algorithm not only is not necessarily the best value but also they can be the worst values. We executed all mentioned algorithms with different values of their main parameters and mentioned the best values for the best result with their best parameter in Table 3.

```

jupyter ML Last Checkpoint: 7 minutes ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)
In [37]: from sklearn.ensemble import HistGradientBoostingClassifier
HGBC = HistGradientBoostingClassifier(learning_rate=0.2, max_iter=500)
HGBC = HGBC.fit(X_train, y_train)
y_predict = HGBC.predict(X_test)
print("accuracy_score = ", accuracy_score(y_test, y_predict))
print("f1_score = ", f1_score(y_test, y_predict))

HGBC_f1_score_arr = np.zeros(100)
HGBC_ind = 0
for num in range (1,100,1):
    HGBC_l_ra = num / 500
    HGBC_m_itr = num * 5
    HGBC_ = HistGradientBoostingClassifier(learning_rate = HGBC_l_ra, max_iter = HGBC_m_itr)
    HGBC_ = HGBC_.fit(X_train, y_train)
    y_predict = HGBC_.predict(X_test)
    HGBC_f1_score_arr[HGBC_ind] = f1_score(y_test, y_predict)
    HGBC_ind = HGBC_ind + 1

print("Max of f1_score: ", HGBC_f1_score_arr.max())
print("Index of Max f1_score: ", np.argmax(HGBC_f1_score_arr))

accuracy_score = 0.8234203041919129
f1_score = 0.8258111734569408
Max of f1_score: 0.8291970802919708
Index of Max f1_score: 47
    
```

Figure. 4. HBGB Classifier Python Code run in Jupyter Notebook

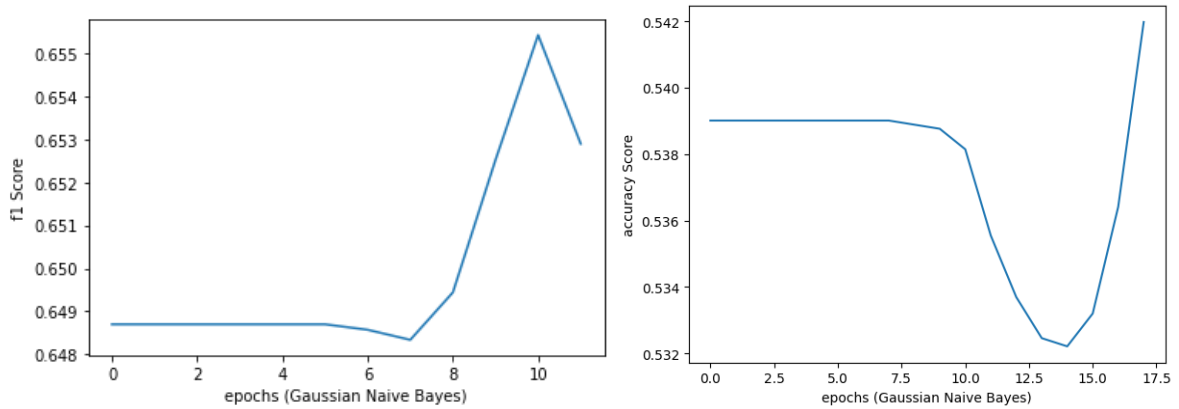


Figure. 5. GNB

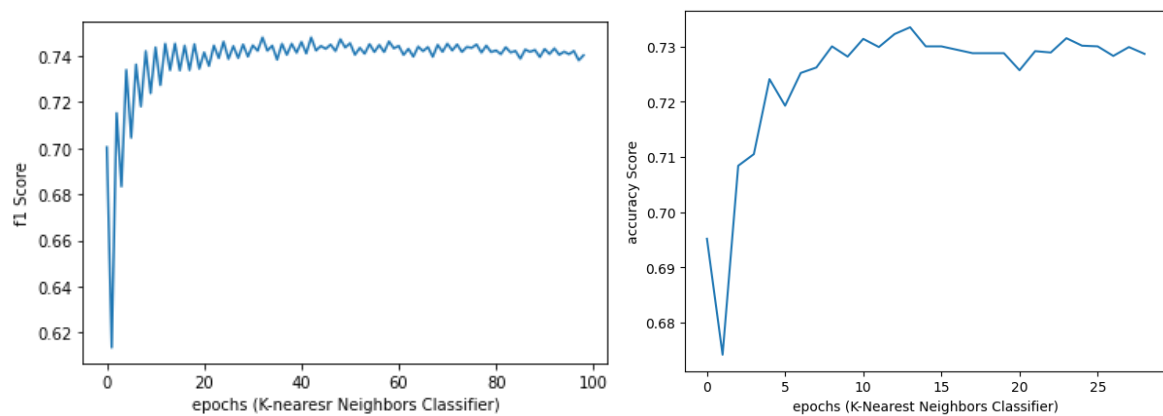


Figure. 6. KNN

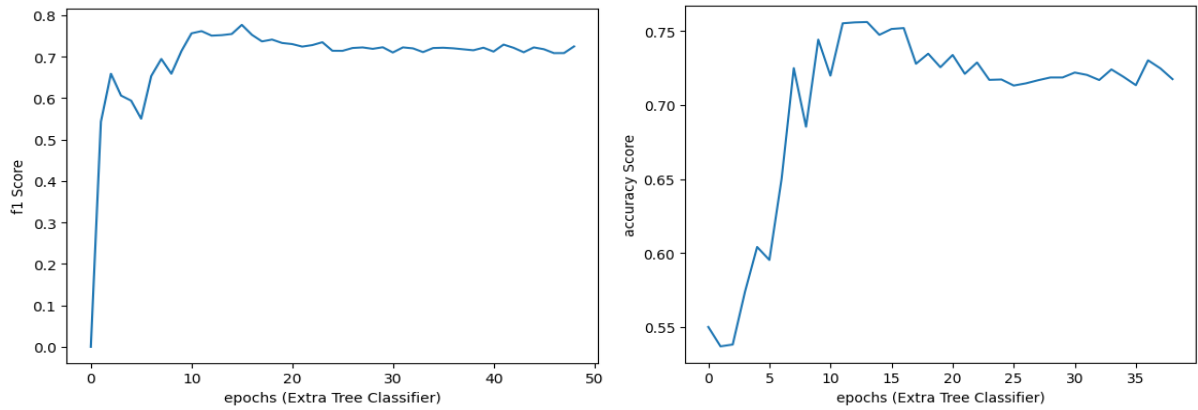


Figure. 7. ETC

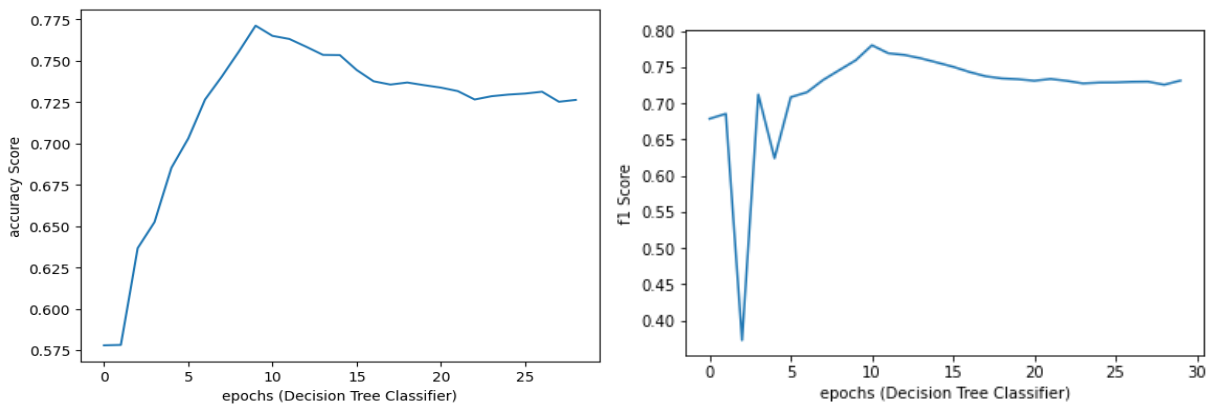


Figure. 8. DTC

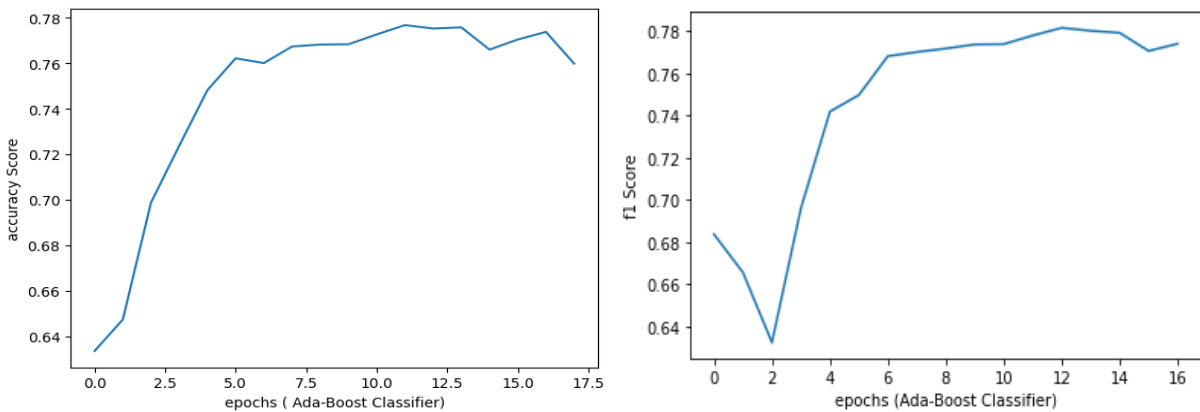


Figure. 9. ABC

In addition, as mentioned earlier, we executed our model in additional epochs with diverse parameters to find optimal amounts for Hyper-parameters to reach better results and accuracy, which we called Hyper-parameter Optimization. We can see the results in Figures 5 to 16 respectively. Because of fast convergence and high execution time, we executed some algorithms in fewer epochs (such as Random Forest).

6. Discussion

In this paper, we tried to analyze the reviews on e-commerce websites with NLP techniques and context-based approaches such as Sentiment Analysis. Then, we designed and proposed a model to distinguish between fake and true ones with Supervised Machine Learning Algorithms. Some key findings were gleaned from this work. We compared our proposed approach with existing approaches

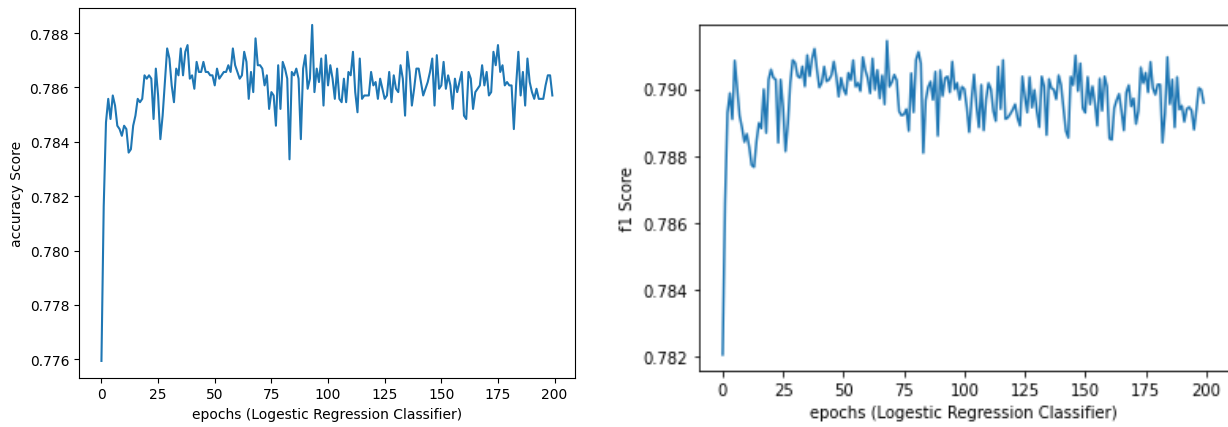


Figure. 10. LR

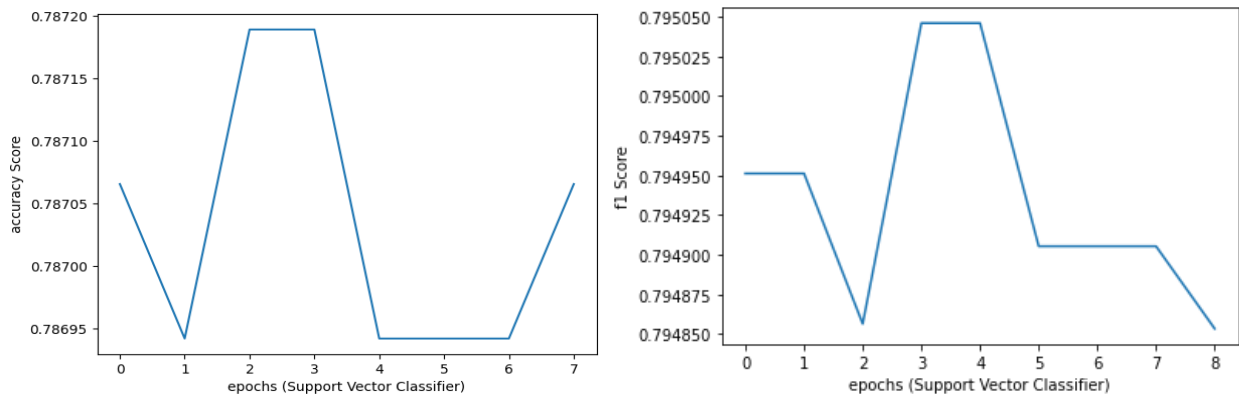


Figure. 11. SVC

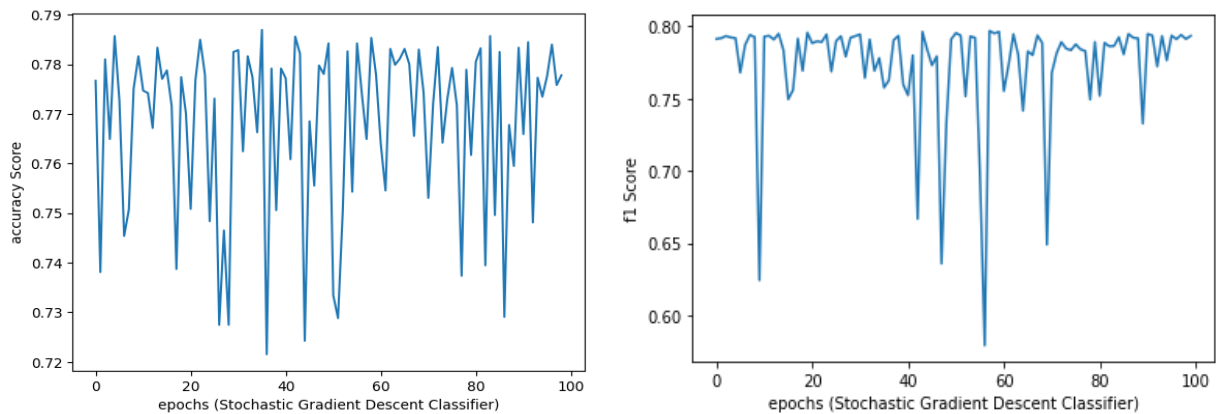


Figure. 12. SGD

and used the F1-Score and Accuracy Score measurement metric to compare and evaluate the model's accuracy. Moreover, our proposed approach was also found to outperform existing approaches, such as [17], [6], [7], [22], [23], [24], [25], and [26] considering metrics measurement. Moreover, the dataset we used in this paper was large enough and formed a significant contribution. The reviews labeled as fake in our dataset were computer-generated reviews that have the

advantage of being undeniably fabricated, as they did not exist before. Also, we have used numerous algorithms, whereas most similar works have used just limited classifiers such as NB, SVM, etc. In this paper, we demonstrated that using a limited number of features for Supervised Machine Learning models is possible. However, prior studies have often attempted to train their model using many features. Moreover, some studies have used n-grams to distinguish between authentic and

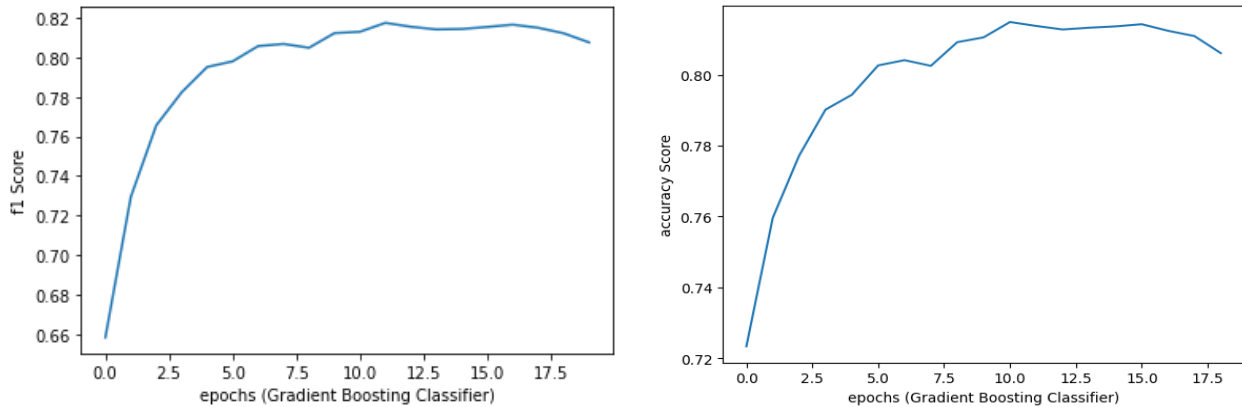


Figure. 13. GB

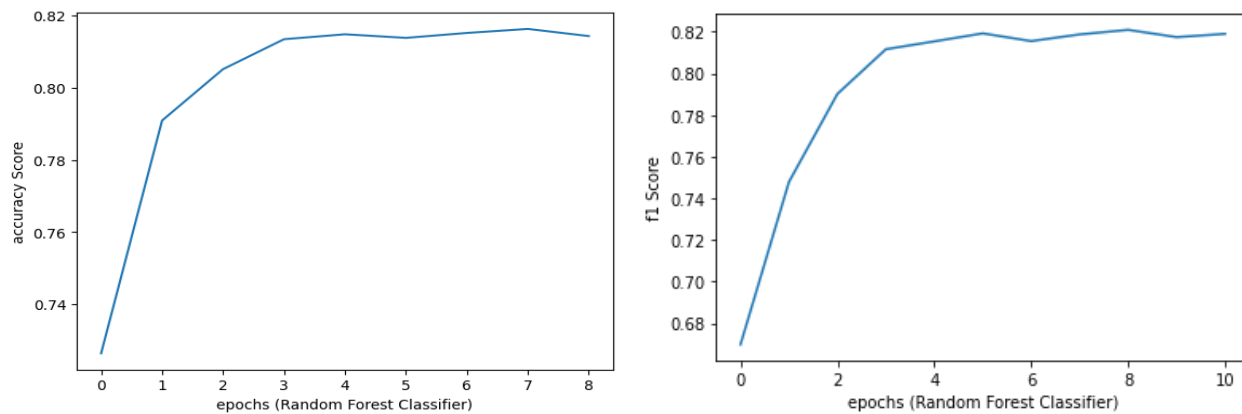


Figure. 14. RF

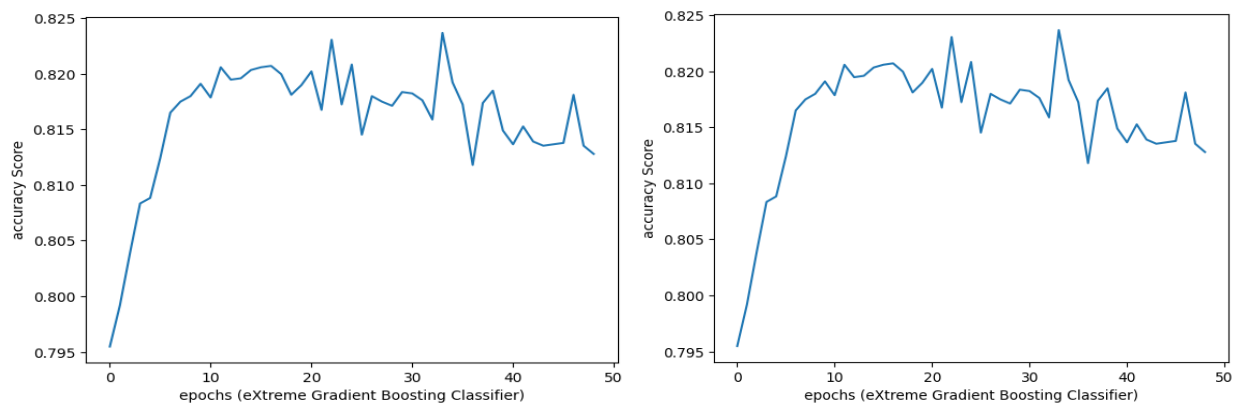


Figure. 15. XGBC

fake reviews. Although such approaches guarantee acceptable performance, such methods are computationally intensive, and the findings could often be merely due to chance. And finally, we showed that the default hyperparameters of Machine Learning classifiers do not necessarily have the best values.

7. Conclusion and Future Work

Thanks to technology and Internet development,

there are tons of online shopping websites. They provide a massive amount of vivid products and services that users continuously comment on them. However, due to the high volume of deception on these websites, not all reviews are trustworthy to be genuine. Nevertheless, users of these websites mainly intend to read all reviews and comments before purchasing when they need to shop. In this research, we tried to reveal the significance of reviews on e-commerce websites and how they

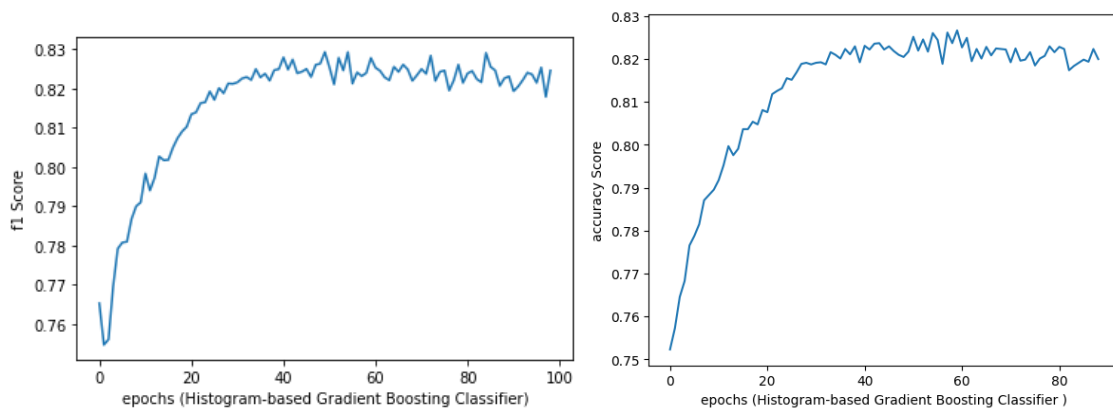


Figure. 16. HBGB

influence almost all users' decisions. In this paper, we did a Sentiment Analysis on review texts to apply better features of the reviews in our dataset. Then a Machine Learning fake reviews detection model was presented. The Amazon reviews and Computer-Generated data were used as a dataset to train and evaluate the proposed approach. Different Machine Learning algorithms and several methodologies are implemented in the developed approach.

We tried to examine the accuracy of all proposed algorithms to find the most acceptable algorithm. Traditional Machine Learning algorithms, ensemble algorithms, and boosting classifiers are used to execute the proposed model. The results revealed that HistGradientBoostingClassifier outperforms the rest of the algorithms in the learning process. It gave the best result among all other classifiers by achieving 82.9% of the F1-Score. This result not only ensured the performance of our model but also suggested the importance of ensemble techniques in comparison with traditional machine Learning techniques. They are more precise than different strategies as they are compelling classifiers and well-suited for 2-class problems. However, this meta-classification algorithm (HistGradientBoostingClassifier), which appeared as the best-performing algorithm in this paper, has not been widely used in related studies on fake reviews detection. On the other hand, our dataset is a mixture of genuine-labeled and fake-labeled (Computer-Generated) reviews. Our approach performs with high accuracy to detect fake and genuine reviews, so this implies that machines can fight machines in the battle against fake reviews.

Reading all positive reviews about a product on websites does not guarantee its high quality. Reviews are valuable just when they are genuine. Our model can be helpful for users to differentiate fake reviews from true ones and ignore them entirely in product purchases. However, detecting fake reviews remains a complex task despite researchers' great efforts in this direction. We should

also acknowledge a few drawbacks of this paper. For one, the content of the examined dataset was limited to Amazon reviews. However, we aim to expand this study using other datasets such as Yelp, eBay, Digikala and trip advisor datasets and use different preprocessing strategies and feature selection methods in future works. Furthermore, we may consider including other behavioral features of the reviewers to develop a reviewer-based approach. It could definitely improve the performance of the fake review detection process. And also we can enhance the accuracy of our model using more sophisticated labeling methods and learning techniques such as Neural Networks and Deep Learning.

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

KNN: Study design, acquisition of data, software design and implementation, interpretation of the results, statistical analysis, drafting the manuscript;

AAPHK: Study design, interpretation of the results, supervision, revision of the manuscript.

Conflict of interest

The authors declare that no conflicts of interest exist.

References

- [1] S. Mani, S. Kumari, A. Jain, and P. Kumar, "Spam review detection using ensemble machine learning," in International Conference on Machine Learning and Data Mining in Pattern Recognition, 2018, pp. 198-209.
- [2] N. A. Patel and R. Patel, "A survey on fake review detection using machine learning techniques," in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1-6.

- [3] J. Salminen, C. Kandpal, A. M. Kamel, S.-g. Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," *Journal of Retailing and Consumer Services*, vol. 64, p. 102771, 2022.
- [4] M. Ennaouri and A. Zellou, "Fake Reviews Detection through Machine learning Algorithms: A Systematic Literature Review," 2022.
- [5] G. Bathla, P. Singh, R. K. Singh, E. Cambria, and R. Tiwari, "Intelligent fake reviews detection based on aspect extraction and analysis using deep learning," *Neural Computing and Applications*, pp. 1-17, 2022.
- [6] S. C. Shetty, "Learning to detect fake online reviews using readability tests and text analytics," Dublin, National College of Ireland, 2019.
- [7] S. Banerjee, A. Y. Chua, and J.-J. Kim, "Using supervised learning to classify authentic and fake online reviews," in *Proceedings of the 9th international conference on Ubiquitous Information Management and Communication*, 2015, pp. 1-7.
- [8] E. Anderson and D. Simester, "Deceptive reviews: the influential tail," *Tech Rep*, vol. 2, p. 1, 2013.
- [9] E. Elmurghi and A. Gherbi, "An empirical study on detecting fake reviews using machine learning techniques," in *2017 seventh international conference on innovative computing technology (INTECH)*, 2017, pp. 107-114.
- [10] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," *arXiv preprint arXiv:1107.4557*, 2011.
- [11] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What yelp fake review filter might be doing?," in *Proceedings of the international AAAI conference on web and social media*, 2013.
- [12] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, 2017, pp. 127-138.
- [13] K. Lee, J. Ham, S.-B. Yang, and C. Koo, "Can you identify fake or authentic reviews? An fsQCA approach," in *Information and communication technologies in tourism 2018*, ed: Springer, 2018, pp. 214-227.
- [14] E. Elmurghi and A. Gherbi, "Detecting fake reviews through sentiment analysis using machine learning techniques," *IARIA/data analytics*, pp. 65-72, 2017.
- [15] H. Li, B. Liu, A. Mukherjee, and J. Shao, "Spotting fake reviews using positive-unlabeled learning," *Computación y Sistemas*, vol. 18, pp. 467-475, 2014.
- [16] S. Noekhah, N. binti Salim, and N. H. Zakaria, "Opinion spam detection: Using multi-iterative graph-based model," *Information Processing & Management*, vol. 57, p. 102140, 2020.
- [17] K. Algotar and A. Bansal, "Detecting Truthful and Useful Consumer Reviews for Products using Opinion Mining," in *EMSASW@ ESWC*, 2018, pp. 63-72.
- [18] N. M. Danish, S. M. Tanzeel, N. Usama, A. Muhammad, A. Martinez-Enriquez, and A. Muhammad, "Intelligent interface for fake product review monitoring and removal," in *2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, 2019, pp. 1-6.
- [19] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 188-197.
- [20] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, pp. 1-167, 2012.
- [21] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *Journal of Big Data*, vol. 2, pp. 1-24, 2015.
- [22] F. H. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," in *Twenty-second international joint conference on artificial intelligence*, 2011.
- [23] P. Hajek, A. Barushka, and M. Munk, "Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining," *Neural Computing and Applications*, vol. 32, pp. 17259-17274, 2020.
- [24] F. Abri, L. F. Gutierrez, A. S. Namin, K. S. Jones, and D. R. Sears, "Fake reviews detection through analysis of linguistic features," *arXiv preprint arXiv:2010.04260*, 2020.
- [25] L. Gutierrez-Espinoza, F. Abri, A. S. Namin, K. S. Jones, and D. R. Sears, "Fake reviews detection through ensemble learning," *arXiv preprint arXiv:2006.07912*, 2020.
- [26] P. K. Jain, R. Pamula, and S. Ansari, "A supervised machine learning approach for the credibility assessment of user-generated content," *Wireless Personal Communications*, vol. 118, pp. 2469-2485, 2021.

Kian Nimghaz Naghsh is a Software Engineer who received his B.S. in Computer Engineering from the Islamic Azad University of Tabriz. He has over 10 years of experience in the field of software engineering. His current research interests include Artificial Intelligence, Machine Learning and Deep Learning.



Ali Asghar Pour Haji Kazem is an Assistant Professor in the Software Engineering Department of Istinye University in Istanbul, Turkey. He received his B.S. in Computer Engineering from the University of Isfahan, his M.S. in Computer Engineering from Shahid Beheshti University, and his Ph.D. in Computer Engineering from IAU Tehran Science and Research Branch. He has published over 40 papers in refereed international journals and conferences. He is the reviewer of different international journals in Elsevier, Springer, and Wiley. His current research interests include Cloud Computing, Edge Computing, Optimization Algorithms, Machine Learning and Deep Learning.

