Sentiment Analysis and Fake Amazon Reviews Classification Using SVM Supervised Machine Learning Model

Myasar Tabany * and Meriem Gueffal

Networks, Security and Systems Research Group, School of Physics, Engineering, and Computer Science,
University of Hertfordshire, AL10 9AB, Hatfield, Hertfordshire, UK
Email: m.tabany@herts.ac.uk (M.T.); mg21aaq@herts.ac.uk (M.G.)
*Corresponding author

***Abstract*—This project attempts to conduct sentiment analysis of short and long Amazon reviews and report their effects on the supervised learning Support Vector Machines (SVM) model, to bridge for fake reviews classification. Firstly, the SVM model was evaluated by comparing its performance against Naive Bayes, Logistic Regression, and Random Forest models and proved to be superior (second assumption) based on the accuracy (70%), precision (63%), recall (70%), and F1-score (62%). Hyperparameter tuning improved the SVM model for sentiment analysis (accuracy of 93%), then altering the review length affected the model's performance, which validated that review length affects the classifier (first assumption). Secondly, conducted fake reviews classification on the fake reviews' dataset yielded 88% accuracy, while the merged subsets of the two datasets yielded 84% accuracy.**

***Keywords*—fake reviews detection, sentiment analysis, natural language processing, Machine Learning (ML) supervised learning**

I. Introduction

The recent pandemic was an example that demonstrated the necessity of online shopping; however, avoiding fraud is still a challenge on e-commerce platforms, such as fake reviews. Amazon stopped more than 200 million deceptive reviews in 2020 and reported 10000 groups to Facebook for engaging in organized review fraud [1, 2]. With this challenge, it is difficult to conduct market research, especially for small businesses that want to use data to apply sentiment analysis to classify positive and negative reviews of genuine customers. Fake reviews can shift their focus on what drives customer satisfaction and predict profit. Fake opinion detection has attracted researchers' attention since 2007 [3–5], and many studies have attempted to employ effective solutions to battle this problem. E-commerce introduces fewer restraints and offers potential for success, which makes it appealing to fraudulent individuals and organizations that want to reach a wider audience quickly and save on marketing expenses through deceptive means [6].

- Ethical and legal issues:

Federal Trade Commission prohibits fake reviews on e-commerce websites [7] and holds accountable companies that allow these fraudulent activities to flourish on their platforms. The Federal Trade Commission (FTC) released a list of companies that received penalties for endorsing deceptive reviews [8], and Amazon was on this list.

- Economic and social issues:

Reviews should reflect the genuine opinion of customers about a product [9]. Sellers and manufacturers use feedback to improve their products [3]; deceptive reviews alter the truth about the quality of the product leading to an increase in product returns and logistic expenses [10]. Therefore, the project focuses on employing the supervised learning model Support Vector Machines (SVM) [11] to study fake reviews depending on their length in Amazon reviews' dataset.

- Professional issue:

The fake reviews detection is not an easy manually, as it is hard to differentiate between a fake review and genuine one by reading [12]. Even with experts' knowledge, it is impossible to accomplish detection with high certainty as spammers learn to adapt to detection algorithms by imitating genuine customers. To draw on how serious the fake reviews phenomenon, it is necessary to mention the extent how difficult spam detection is on e-commerce that for instance, Yelp resulted to public shaming businesses that resolve to fake reviews to prosper their sales [12]. This study is motivated by investigating reviews in various word counts to deduct characteristics of fake Amazon reviews which raises this question: Do long reviews affect the supervised learning model performance?

The assumption that the longer the review the higher the probability is a trustworthy one, this stems from the fact that fake reviews do not normally contain a strong sentiment about the product. Genuine reviewers post high-quality feedback that is rich in useful information and extensive description of the advantages and

Manuscript received February 20, 2023; revised April 10, 2023; accepted June 5, 2023; published January 9, 2024.

doi: 10.12720/jait.15.1.49-58

drawbacks of the product reviewed [13]. To investigate this further, determining a high performant classifier is needed to analyze the reviews. For this purpose, a selection of supervised learning models is compared before performing the classification of positive and negative reviews in the Amazon dataset. This raises another question: Is the Support Vector Machines (SVM) model suited to apply sentiment analysis and fake reviews detection of Amazon customer reviews. The assumption is that the Support Vector Machines model is superior to other supervised learning classifiers in analyzing e-commerce reviews. From the literature, different machine learning approaches were employed that involved supervised, semi-supervised, unsupervised, and deep learning methods [14].

Our main research contribution in this paper is proposing a robust model to classify positive and negative reviews and fake and genuine reviews using sentiment analysis and opinion spam detection by employing word count as a determinant. Supervised models serve different purposes and offer several advantages; to choose the optimal one, it is necessary to compare popular supervised models from the literature against the Support Vector Machines (SVM) and then improve their performance to achieve highly accurate classification of the target labels. Since reviews can range from no words to very long ones, it is unclear what word count range fake reviews fall into.

On the other side, the research study contributes to the investigation of the average review length of fake Amazon reviews. This project's objectives are to employ several advanced supervised machine learning algorithms and their use cases based on their strengths and weaknesses, and to discover the current practices in both sentiment analysis and fake review detection. This has resulted in implementing the Support Vector Machines (SVM) and demonstrating their effectiveness in analyzing e-commerce customers' reviews.

The remainder of the paper is structured as follows. Section II reviews the highly pertinent literature to the paper's topic while highlighting related works. The experimental side, which included the main intensive implementation, materials, and methods used in the practical work suggested for the model, were thoroughly explained in Section III. Section IV discusses the study findings and its analysis, and finally Section V includes a conclusion and suggestions for future work.

II. Literature Review

This section basically supports the background section by providing evidence for the proposed hypothesis. This section should be more comprehensive and thoroughly describe.

Ya *et al.* [15] proposed two methods for detecting spam reviews using the Topic Model and Reviewer Anomaly and Reviewer Anomaly which extract topics combination from the review content. The Linear Discriminant Analysis (LDA) model delivers a high probability of top 6 features. The authors calculated the reviewer's abnormality degree based on the extracted features and their weight, time, and similarity. A set threshold of high and low scores for each review determined the spam reviews. The authors claimed these models were effective and have improved fake reviews detection by evaluating their precision, recall and F values.

Zhang *et al.* [16] proposed an unsupervised shape let model to group stores based on the time series and the similarity of the review features to build a heterogeneous network linking stores, products, and users. The authors observed that some sellers employed fake reviewers who gradually increased their activity to avoid getting detected by Taobao's algorithms. Since fake reviewers write reviews in exchange for a bonus, most form groups to increase profit, many e-commerce platforms limit this behavior by allowing only purchasers to provide product reviews.

Xiang *et al.* [17] proposed the Convolutional Neural Network (CNN) to detect fake reviews by combining behavioral features in users and text features in products to extract temporal features and train a Long Short-Term Memory (LSTM) model. The classification was performed by inputting vectors representing the features into the classifier and then measuring its performance by accuracy, precision, recall, and F1-Score. The authors observed that the early reviews control the sentiment around the product and gain users' attention by giving higher ratings to products to maintain a top position. In addition, most spammers write more reviews daily than genuine users.

Hussain *et al.* [18] used two techniques to detect fake reviews, the Behavioral Method (SRD BM), which relies on behavioral features such as time and ratings to measure a spam score, and the Linguistic Method (SRD-LM), which relies on content-based features such as text to classify reviews. Both models performed well based on evaluating their accuracy, especially SRD BM because it deeply analyses spammers' behavior. The authors then used SRD BM to produce a labelled dataset and used SRD-LM to train and compare performances of Naïve Bayes, Logistic Regression, Support Vector Machines and Random Forest classifiers. SVM with unigram and Information Gain (IG) achieved high performance while Logistic Regression was deemed superior.

Rout *et al.* [19] applied semi-supervised learning to detect spam in hotel reviews since supervised methods require quality data to assure superior performance. The authors relied on linguistic features and the Positive Unlabeled (PU) learning-based classification that improved the F-score metric. Ren and Ji [14] analyzed and compared statistical and deep learning models employed to detect fake reviews and then summarized their findings. The authors discussed the unresolved issues related to building datasets and algorithms' adaptation to domains. Studies used crowdsourced data to create datasets because labelling is challenging, and their evaluation does not reflect detection in real fake review cases.

The authors categorized building datasets methods into:

- Rule-based methods where the researchers applied the general rules observed from the patterns in the data, such as duplicated instances and similar content.

- Human-based methods where researchers employed human judges to annotate the values and then agree on meeting a decision score.
- Filtering algorithms where researchers experimented with a set of features and approaches to imitate a highly confidential and reliable system.
- ATM-based methods: researchers employed Amazon mechanical Turks or crowdsourcing services to generate a large-scale synthetic fake reviews dataset.

They reported that evaluation in balanced datasets by accuracy, precision, recall and F1 score. If unbalanced, the Receiver Operating Characteristics (ROC) curve and Area Under the Curve (AUC). Some studies relied on text or behavior features or a combination of both. They outlined popular models in supervised learning: SVM, Naive Bayes, Logistic Regression, SAGM, Semi-supervised: co-training, Positive Unlabeled (PU) learning, Unsupervised learning: Semantic Language Model (SLM), and Neural networks: CNN, RNN, and others.

Zeng *et al.* [20] proposed an approach to select deceptive from truthful reviews by analyzing the sentiment in the first, the middle and the last sentences separately using four bidirectional Long Short-Term Memory (LSTM) models, then treated the outputs with self-attention mechanism layers and classified the results with a fully connected neural network. They observed that the first and the last review sentence have strong sentiments compared to the middle one, and fake reviews start and end with similar sentences and show stronger opinions than truthful reviews.

Zhang *et al.* [13] proposed the Co-training for Spam (Copa) review identification in an unlabeled hotel dataset employing two views to analyze text-based features and syntax using Probabilistic Context-Free Grammars (PCFG) rules. The authors found that deceptive reviewers used past tense in writing imaginary information, while truthful reviewers used present tense with noun phrases to reflect their real detailed and specific experiences. Zhang *et al.* [21] proposed the Co-training by Features (Cuphea) semi-supervised model to detect spam in unlabeled reviews dataset using two views that rely on lexical terms sorted by entropy score and SVM as a base classifier. Entropy is the size of the textual content for a review depending on its uncertainty, using this score to form two subsets based on the number of term lexicons. Kale *et al.* [22] investigated characteristics such as text flow, use of offensive language, and out-of-context reviews that help identify fake reviews, then checked for similarities in customer reviews to build a graph to illustrate the relationships.

Ott *et al.* [23] developed three methods to detect deceptive opinions and created a dataset of 400 reviews for 20 hotels using Tuckers to compose deceptive reviews and three undergraduate students to judge the truthfulness of a subset of it. The authors relied on linguistics and psychological cues to detect lies, such as negative sentiment and mental distancing and used n-gram and linear SVM as classifiers for similarities based on POS (part of speech) distribution. They reported that truthful reviews were sensitive and informative while deceptive ones were more imaginative and focused on external descriptions and ensured presence increasing first-person use. Kangal *et al.* [24] compared different supervised and semi-supervised approaches used to detect fake reviews and authentic reviewers by accuracy, precision and recall applied to raw and pre-processed data. The authors reported that Logistic Regression was superior, followed by random forest that performed better than SVM.

Ott *et al.* [6] proposed the SVM model to classify the reviews and the Naive and the Bayesian Prevalence Models with Gibbs sampling to estimate the deception rate based on economic signaling theory cues such as posting requirements and exposure rate in positive hotel reviews. They noticed deception rate decreased by filtering first-time reviewers from the data, and what influence deceptive reviews is purchasing the product to review and the audience size that will read feedback, fake reviews are highly prevalent on platforms with no posting requirements and exposure.

Taqiuddin *et al.* [25] used lexicon-based features: sentiment, content, for instance, the number of likes on reviews, Metadata considering the length of text, and Profile time spent by reviewers on the platform, then Term Frequency-Inverse Document Frequency (TF-TDF) weighting to classify deceptive Steam reviews using SVM, assessing to the accuracy, precision, recall, and f-measure metrics and deployed on a dashboard for stream users to make informed purchase decisions.

Wang and Zhu [26] used the dataset from Ott *et al.* [23] with linear SVM to classify reviews and then a voting scheme for feature weights to improve the detecting fake reviews. To reduce the text feature dimensions, they compared the pre-processing techniques n-gram, POS-tag and TF-IDF in textual feature and used Latent Semantic Indexing (LSI) and latent semantic analysis (Sprinkle). They found that spam reviews have excessive use of 2nd person pronouns and less concrete nouns and objectives. Lighters *et al.* [27] compared semi-supervised approaches' effectiveness in classifying opinion spam in unlabeled data based on the review content since graph-based semi-supervised learning relieves the effort and cost of labelling data when only a small number of labelled data and a large number of unlabeled data sets are available. They reported that self-training with naive Bayes with TF-IDF top 1000 features performed better on the golden standard dataset against supervised models while performing similarly to supervised models with Naive Bayes on the Yelp dataset. The authors also found that the supervised model SVM performed better with 10% labelled data than 20%, and the models performed effectively with one type of polarity, explaining that the data had similar instances.

Patil *et al.* [28] proposed a method to detect fraudulent reviewers' behavior and interactions and remove fake reviews based on IP address and user ID similarities. They used the bag of words method to indicate if a review is positive or negative, then TF-IDF to organize the word relevance for SVM to classify the review as truthful or fake.

Padma *et al.* [29] achieved high accuracy using words' sentiment extraction to detect spam reviews on the reviewer level. They compared Naive Bayes, SVM and their proposed model. They noticed the models' improvement depends on reducing misidentifying labels and time complexity. Ye *et al.* [30] used a temporal method to detect fake reviews by tracking linguistic, relational, and behavior signals over time for abnormal occurrences (bursts of activity) in real-time to manually examine factors of spam reviews. The authors highlighted the importance of the spam activity timeline, focusing on detecting spammers' complicit promotion or demotion activities and sudden changes in the average rating over time using the number of positive and negative reviews and observing sudden increases that may affect the overall rating.

Khurshid *et al.* [31] evaluated the Ensemble Learning Model (ELM) a set of models as base classifiers for spam detection in reviews by the precision, recall, f score and Receiver Operating Characteristic (Roch) metrics. The authors applied the Chi-Squared technique for feature extraction and selection to exclude duplicative and unimportant features to build a high-performant system with low cost and time consumption. They compared the Extreme Learning Machine (ELM) system against Multi-Layer Perceptron (MLP), Naive Bayes, and Ad boost using all the features, then with a set of features selected by the chi-squared technique to evaluate their impact on the system proving to be robust. The limitation of their study is the dataset imbalance where spam reviews are less than genuine ones and the lack of large-scale spam review datasets for supervised learning applications.

Savage *et al.* [32] observed the average rating of the products reviewed and compared it to the mean rating of honest reviewers using binomial regression to identify characteristics of reviewers' behavior engaging in spam reviews by deviating from the overall rating in existent and synthetic datasets. Their approach tried to avoid the disadvantages of review-centric methods that focus on text features to identify duplicate reviews as spam which become impractical when comparing more data or manual classification, relying only on ratings since leaving a review is optional. The approach showed performance improvement and potential for effectiveness with other features. Li *et al.* [33] proposed a two-mode Labelled Hidden Markov Model based on linguistic and behavioral features for spammers and spammer groups detection in hotel reviews using review time and compared its performance with supervised learning models based on the Accuracy, Precision, Recall and F1-score on five-fold cross-validation. Dianping's filtered and unfiltered reviews enabled the authors to discover the users' bimodal temporal patterns and to define two posting states to differentiate between spammers (active/fast rate) and genuine users (inactive/slow). The authors' analysis showed that the spammers wrote fake reviews consecutively in a short time compared to other reviewers, detecting relations between group spammers with raised accounts. Shehnepoor *et al.* [34] proposed the NetSpam framework to model reviews as a Heterogeneous Information Network (HIN) for detecting fake reviews using spam features weighted based on their importance in the Yelp and Amazon datasets. The results showed that their approach performed better than other methods using fewer features, especially review-behavioral features, and can reduce computation costs.

Long *et al.* [35] categorized review types into non-reviews consisting of no opinion or advertisement, brand-only reviews that evaluate the brand but not the product, untruthful reviews that deliberately deceive users, and off-topic reviews. The authors reported that spam detection in the non-reviews category achieved the highest performance, and the helpful reviews feature might help detect spam. Gupta *et al.* [36] created a feature-based model to select extremist reviewer groups and manually labelled 923 Amazon reviewer groups by three annotators who considered the length of the reviews, their similarity, use of capitalization, and brand advertising in the title, text, date of posting and number of helpful votes for each product. The authors compared supervised classifiers by the precision, recall, F1 score and Area under the ROC curve (ROC-AUC). The neural network 3-layer Perceptron performed the highest while the Decision tree performed poorly. The authors observed that fake verified reviews are possible since spammers get refunded for their purchase, the reviews number about a brand strongly indicates extremism, and extremist groups write more 5-star reviews than other users. Fazzolari *et al.* [37] observed the Cumulative Relative Frequency Distribution, the occurrence of the value of features effectively used for spam detection in previous studies on the labelled Yelp dataset. They relied on review-centric features such as the number of pictures, votes, Average Gap and Average Rating Deviation and review level features such as reviewer expertise, reviewer activity, and first review. The authors recorded that the performance of the supervised approaches achieved superior performance with Naive Bayes and Support Vector Machines as models.

Salunkhe [38] proposed an Attention-based Bidirectional LSTM model to classify truthful and deceptive reviews on a balanced dataset using semantic content features and compared supervised and deep learning classifiers. In supervised learning, Multinomial Naïve Bayes followed by Logistic Regression performed better. Deep Learning models performed better. Mukhrejee [39] proposed an approach to classify negative opinions in product reviews using the Amazon dataset. The author stated that positive sentiment is easier to detect, whereas negative sentiment in product reviews contains issues in specific product categories phrase extraction using Max Ent Aspect Sentiment Model (ME-ASM). Li *et al.* [40] relied on Dianping's filtering algorithm to build a large-scale dataset of fake and authentic restaurant reviews since the algorithm considers a reviewer spammer if more than half of their feedback is labelled fake. The authors relied on features such as IP addresses and user profiles to find patterns. They discovered that spammers begin their activity as soon as they join the platform resulting in a high fake reviews volume. In addition, spammers conduct their campaigns every day except for Mondays, but other users post on Sundays. The authors employed the Average Travel Speed (ATS) metric to measure the abnormal mobility rate. The authors assumed that professional spammers exhibit when changing their IP constantly to register on new accounts in a short period to bypass the detection system, which explains the abnormal changes in their location. They combined spatial and temporal with n-gram and behavioral features with the SVM model and showed high efficiency in opinion spam detection.

Jindal and Liu [4] investigated fake reviews using logistic regression's probabilistic output and a combination of features that gave better results. They categorized feedback types into fake opinions, reviews on brand only and non-reviews (advertisements and no opinion reviews). Because there is no labelled dataset, the authors used duplicate and similar reviews to build the spam detection model. Although the authors found helpful feedback was independent of the review's classification, top-ranked reviewers were less trustworthy than bottom-ranked reviewers because they write more reviews and often deviate from the average rating. As spam detection methods advance, spammers find innovative approaches to adapt to the strict measures imposed by e-commerce platforms. Fei *et al.* [5] investigate the change of authorship spamming phenomenon, where spammers seek to buy and sell reputable accounts that exhibit normal history reviewer behavior. The authors employed the single change point detection algorithmic approach Changed-Hands Accounts Detection (CHAD) to statistically identify the time point of the change in writing style and content in these accounts, the CHAD algorithm detects review similarities to identify a single reviewer in different accounts and differentiate the same user from other users.

Deducting from the studies presented in the literature review and the related work sections, a summary of the few points highlighting the research gap:

- Large-scale labelled datasets are not publicly available to research fake review detection, and crowdsourced or human-annotated data cannot substitute for this lack.
- The need for additional information about the users, such as location, IP and MAC addresses, and time spent writing reviews.
- Fake review filtering systems used by e-commerce platforms are confidential; therefore, researchers cannot utilize this knowledge.

III. Materials and Methods

A. Development Environment Specifications

Visualization libraries: matplotlib plotting library version 3.3.4, and Seaborn graphics library version 0.11.2. Natural language processing libraries: Natural Language Toolkit (NLTK) version 3.7, and Regular Expression (RE or RegEx) version 2022.7.25 and Machine learning libraries: Scikit-learn [1] version 1.1.1, Panda's version 1.4.3 were used in the research study.

B. Data Overview, Visualization Analyisis and Preprocessing

The first dataset used for sentiment analysis is the Amazon customer review dataset provided by Amazon S3, containing 130 million customers available for academic research, and comprising of 46 subsets corresponding to each category, in addition to multilingual subsets from outside the US. This project used English written data in the Electronics category based in the US market. The reason for choosing the electronics category dataset was based on the intuition that some products are more susceptible to fake reviews than others. Low-quality products have high rate of fake reviews [4]. To reduce the computation time and effort, only a subset of the dataset 10,000 to reduce problems with loading the dataset the pre-processing it. In early experiments, the dataset was too slow to perform natural language processing. Before starting the machine learning task, it is important to familiarize with the data so observe any noticeable trends or patterns. Since the first part of the project involves sentiment analysis of positive and negative reviews and report the findings. By exploring the dataset categories and generating plots for the star rating distribution. It is observed that most reviews bear a 5-star rating which agrees with Jindal and Lui's findings [4]. Example of star rating plot of the Apparel category are depicted in Figs. 1 and 2.

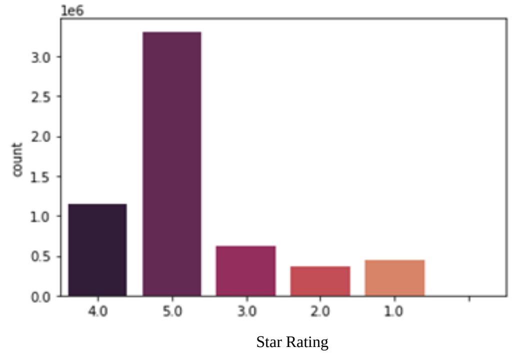

Fig. 1. Distribution of reviews by star rating in the apparel category.

Another example of the Automotive category highlights how the star ratings contain other value types such as dates due to mistakes.

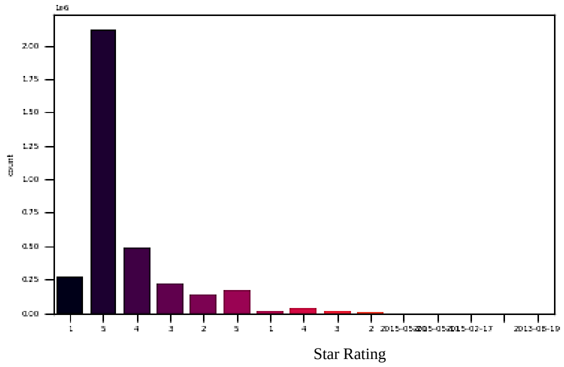

Fig. 2. Distribution of reviews by star rating in the automotive category.

Another useful information observed in the data is the review date, which is useful to learn about the review posting patterns and periods when the reviewing activity increases. The example below in Fig. 3 shows the reviews per month in the home entertainment category.

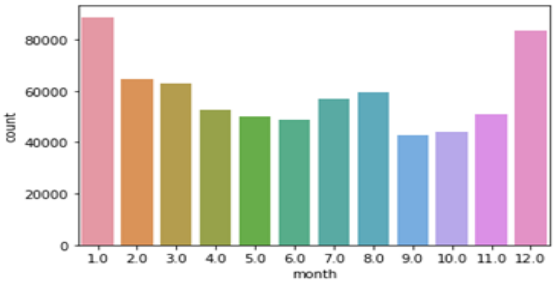

Fig. 3. Reviews per-month distribution in the home entertainment category.

The plots can explain purchase patterns and the popularity of these products (due to the y-axis count), it can be deduced that holidays impact the behavior of Amazon customers. In the home entertainment category, it can be noticed that the activity increases around Christmas and new year times (December-January). This information can help in fake reviews detection, for instance to focus on the time periods of high activity where deceptive reviewers might belong. Since this study's focus is the review word count. Each category is analyzed by number of rows, minimum maximum and average word count, the start and end date for when the data was input, or categories existed. Table I shows information of some categories in the dataset:

TABLE I. VOLUME, MINIMUM, MAXIMUM AND AVERAGE REVIEW LENGTHS OF DIFFERENT AMAZON PRODUCTS/CATEGORIES

Category	**Volume**	**Min Review Length**	**Max Review Length**	**Average Review Length**
Books	3105370	0	10,730	152.96
Gift card	148310	0	1704	24.20
Digital software	101836	1	5228	67.23
Electronics	3091024	0	8360	68.96

From Table I, it can be noticed that most of the minimum review length is 0. The reason is that Amazon does not mandate for customers feedback to have text as they are optional, so users only leave star ratings. The minimum average review length on this dataset is 24 in the gift card category, this can be explained by the nature of these products which can be digital or physical and only contain personal information but nothing else to describe other that the functionality advantages or issues.

IV. RESULT AND DISCUSSION

The maximum average review length is the books category, particularly the hard-cover products to not confuse it with the eBooks category. This word count is explained again by the nature of these products, a book contains information and can be described and talked about in many ways: a user can elaborate on the useful information in the content or the captivating story or can feel inclined to recommend what they found interesting and entertaining about the book, the writer, the genre, or the publisher. Most reviews on this dataset range between 40 and 49 words in length, please see Fig. 4 for more details. More in-depth analysis is needed to analyze word count in positive and negative reviews separately. Example of positive and negative reviews word count: Example of review length per 1-star and 5-star ratings in the music category.

From the dataset with the example above included, it is noticeable that positive reviews (5 star) are longer than negative reviews (1 star). This is because the 5-star positive reviews make most of the reviews on Amazon with them representing more than 50% [4], the users' behaviors sharing their positive experiences is different when they express negative sentiment about the product [33]. It is natural for customers to praise a product if they feel good about it and avoid it if they feel bad, which may affect the way they express their feedback.

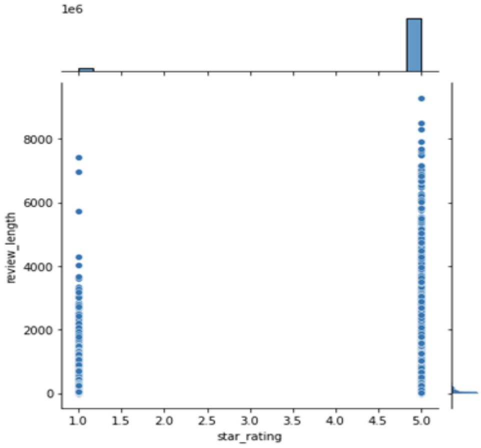

Fig. 4. Distribution of star rating in the review length/ music category.

Conducting in depth analysis is key to learning about the dataset for the following steps before conducting the experiments to verify the assumptions. Data pre-processing is the most critical step in machine learning, it prepares the data for the training and enables a format that is easy to read for the model. The Natural Language Processing (NLP) techniques are applied before proceeding with building and training the algorithms on the dataset:

Lowercase: turning the review body column values in all lowercase not only to make the reviews consistent but also to reduce features since the algorithm does not differentiate between for instance the inputs "good", "GOOD", and "Good" instead it treats them as different values. Punctuation: removing punctuation to make the text readable, removing punctuation makes the next steps of transforming the data easier as it drops unnecessary information in the data and reduces the number of features. Tokenization: it breaks the text into small pieces called tokens, the text becomes easier to the machine to understand and process. Stop words: stop words are words that do not add or subtract from the sentiment of a sentence, removing them from the reviews to decrease the features. Examples of stop words: “the”, “is”, “and”.

Lemmatization: a dictionary-based library that reduces words to their root forms. For example: “Shared” to “share” [40]. Label encoding: encoding the y variables with values between 0 and the number of classes minus 1 [1]. Vectorization: transforming all the input data instances into a feature matrix [1]. The label encoding and vectorization steps are to be executed after splitting the dataset into subsets where the label encoder was applied the target variables while the vectorization was applied to the input variables.

A. Sentiment Analysis

To evaluate the assumption that review length affect the model’s performance. It is needed to compare a few supervised models against the desired one. To validate whether Support Vector Machine (SVM) model it is appropriate to conduct the sentiment analysis on reviews, it is necessary to compare it to other supervised learning methods. In this instance, the SVM model was compared to Naïve Bayes, Logistic Regression, and Random Forest. All these models were used in their default settings.

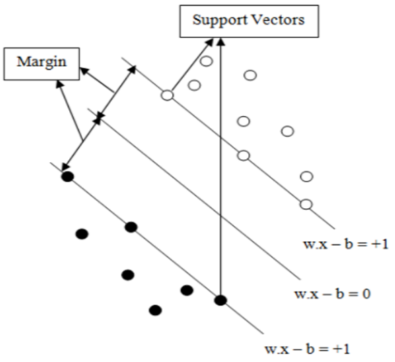

Fig. 5. Overview of the SVM model representation.

- **SVM:** Support Vector Machines (SVM) shown in Fig. 5 [11] is a supervised classification method that uses a hyperplane to sperate classes (Salunkhe, 2021), the aim is to form a large margin that separates the data points called support vectors [21, 24, 39].
- **Naive Bayes:** Naïve Bayes is a method that calculates the probability of a class relying on its features based on Bayes’ theorem it assumes that all features are independent [18, 38].
- **Logistic Regression:** Logistic regression is a method that computes the outcome of a feature [24] by giving a probability that it belongs to a binary class [18].
- **Random forest:** Random Forest is a method based on average estimation of decision trees [18]. In Scikit learn documentation [1], the definition of each metric is provided. To measure the performance of all the compared models.

Accuracy = True Positive + True Negative/All the Values (1)

Precision = True Positive/True Positive + False Positive (2)

Recall = True Positive/True Positive + False Negative (3)

F1-Score=2×(Precision×Recall) / (Precision+Recall) Negative(4)

Salunkhe [38] explained that the accuracy is a measure to assess the model’s performance to accurately classify instances, it calculates the correct predictions out of the total number of observed values. Relying on this metric sometimes is not enough to assess the effectiveness of the machine learning approach so precision, recall and F1-score are good criteria. Precision is the number of the target class predicted correctly out of the total predictions of the same class; recall is the number of correctly predicted values out of the total number of correct predictions. The F1-score is the average of the precision and recall.

TABLE II. THE SVM, NAÏVE BAYES, LOGISTIC REGRESSION AND RANDOM FOREST PERFORMANCE METRICS

Machine Learning (ML) Technique	Accuracy	Precision	Recall	F1-score
SVM	0.70	0.63	0.70	0.62
Naïve Bayes	0.23	0.42	0.23	0.29
Logistic regression	0.69	0.61	0.69	0.62
Random forest	0.66	0.57	0.66	0.56

From Table II, in terms of accuracy SVM performed better than the rest of the models. Therefore, it will be the chosen model for the classification tasks. The SVM performance however needs to be improved, grid search was employed to find the best parameters [1]. After prepossessing the data and preparing it for training, it can be noticed that the model has improved achieving 93%.

To observe the effect of review length on the model’s performance, the sentiment classification code was executed on different ranges of the review word count and recorded the observations separately.

TABLE III. THE SVM MODEL’S EVALUATION METRICS BY REVIEWS WITH DIFFERENT LENGTHS

Review length	Short (≤35)	Long (>35)	> 100	>200	<20	<10
Accuracy %	93	91	88	84	95	95
Precision %	93	91	87	71	94	95
Recall %	93	91	88	84	95	95
F1-score %	93	91	86	77	94	95

From Table III, it can be noticed that as the review length increased the metrics decreased, noticeably when the word count passes the average length of the reviews. For an imbalanced dataset, the weighted values are considered because they reflect each positive and negative class ratio in the target. The confusion matrix [1] is another classification evaluation method, it offers a further breakdown of the algorithms behavior making the correct and false predictions out of the positive and the negative classes.

The results depict how the model is affected by the review length in distinguishing positive and negative reviews for sentiment analysis. When the review is longer the feature set is larger which means more sentiment information input for the classifier to make a prediction.

Decreasing metrics values, especially the accuracy, means that the effort grows to make a distinction between a positive review and a negative one. Since the ratio of 1-star rated reviews to 5-star reviews is low, the dataset is unbalanced so the model is in favor of predicting positive reviews easier. As far as the classification of fake and truthful reviews, there is no hard evidence that all short reviews are fake. Similarly, it is possible that longer reviews contain fake reviews. The sentiment analysis conducted confirms that the SVM model behaves differently according to the length of the reviews, it also indicates that it would be easier for the model to detect fake reviews that are shorter in the word count if run on a labelled dataset. In this case, the dataset at hand does not bare fake and genuine reviews. Several studies resolved to manual selection to overcome this limitation; manual annotation is still less effective than machine learning approaches.

B. *Fake Reviews Classification*

In face of abundant unlabeled e-commerce data, labelled deceptive and genuine reviews datasets are almost non-existent. Many studies have claimed to construct datasets that will be publicly available however, it was impossible task to find such data.

Dataset: Salminen *et al*. [41] worked around the disadvantages of crowdsourcing fake reviews compared to the fake reviews on the Amazon website. The authors artificially generated fake reviews using the language model GPT-2 from the Amazon dataset then evaluated their approach using annotators to detect the newly generated fake reviews in the dataset. The built dataset is publicly available at “https://osf.io/tyue9/?view only= “and comprises 40k product reviews that are split evenly between fake and real Amazon reviews, and has 10 categories: Home and kitchen, sports and outdoors, Electronics, movies and TV, tools and home improvement, pet supplies, Kindle store, books, toys and games, clothing shoes and Jewelry.

The dataset contains 4 columns:

- **Category:** the product category
- **Rating:** a scale from 1 to 5 stars with 1 being the lowest.
- **Label:** either OG or CG. OG stands for original review; the review is assumed to be authentically written by a user. CG stands for a computer-generated review that is fake.
- **Text:** The content of the review

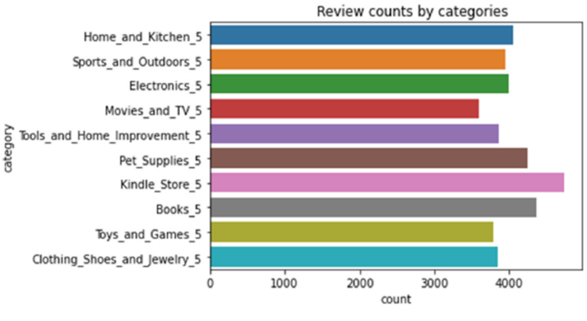

Fig. 6. The reviews in each category in the spam dataset [41].

As the Fig. 6 above shows the number of reviews per category in the dataset, different to the first dataset that had one category. By visualizing and analyzing the dataset (Appendix E), the labelled dataset preserved the same star rating distribution of the Amazon review dataset used in the sentiment analysis experiment as shown in Fig. 7.

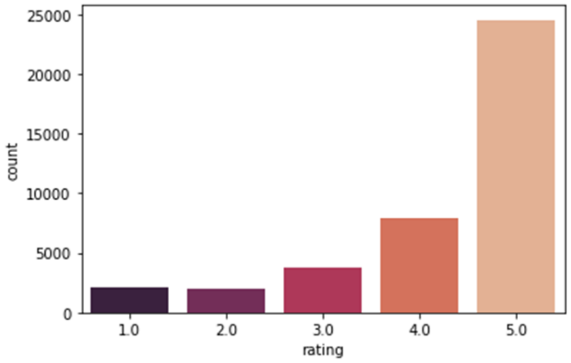

Fig. 7. Star rating distribution in the fake reviews’ dataset.

From Figs. 8 and 9, the review length in the original reviews is longer than the fake reviews. The average review length in the original reviews is 54.75 while in the fake reviews is 46.30. Conducting the fake reviews classification using SVM yielded 88% in accuracy which was less than the accuracy in the sentiment analysis of the Amazon reviews’ dataset (93%).

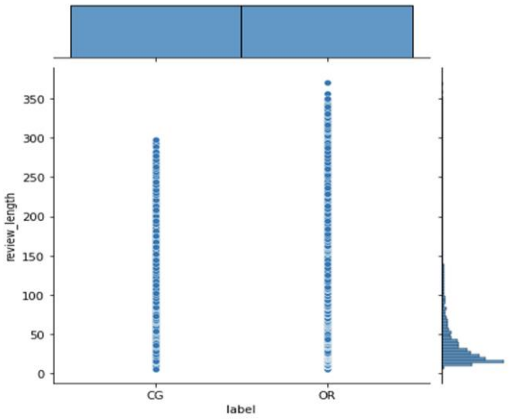

Fig. 8. Review length by label distribution in the fake reviews’ dataset.

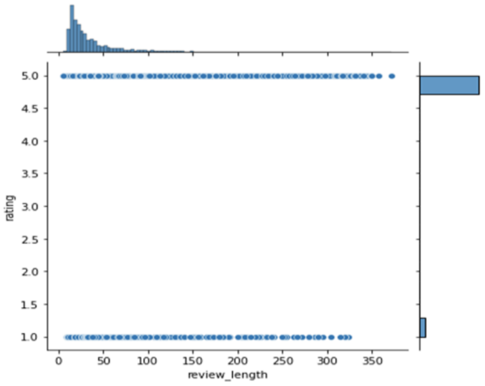

Fig. 9. Review length in the rating distribution.

C. Merging the Datasets

In an attempt of to observe the effects of unlabeled data, this experiment aimed to show the behavior of the SVM model and supervised learning in general. Two subsets of the two datasets were merged to form a new one, taking 10% of the Amazon reviews dataset and the rest from the fake reviews' dataset (10,002 rows in total) to conduct fake reviews classification.

The model's performance yielded 83% in accuracy, less than the result of the fake review classification experiment which may explain the SVM model's behavior in dealing unlabeled instances in the data. Fig. 9 shows the values of the true negative and positive classes, and false negative and positive classes in both predicted and actual values. It shows that adding the unlabeled instances from the first dataset did not significantly affect the model which confirms that the supervised learning approaches are not suited when labelled data is not possible to attain. A visualization of the confusion matrix is depicted in Fig. 10.

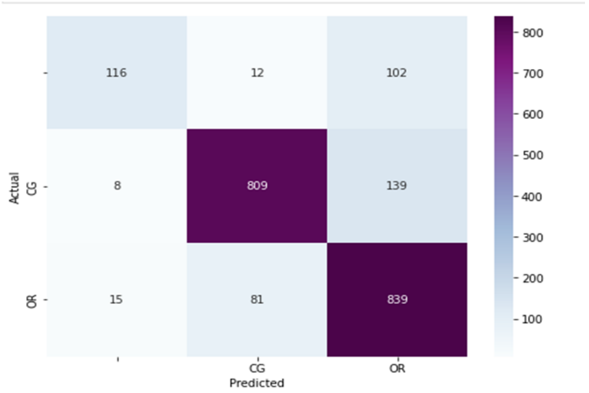

Fig. 10. Visualization of the confusion matrix.

In the first experiment, a selection of supervised learning models' comparison was conducted. The Support Vector Machines model proved to be superior (70% accuracy) against Naïve Bayes (23%), Logistic Regression (69%), and Random Forest (66%). Using the SVM model for sentiment analysis on the first dataset achieved better results after hyper-parameter tuning [1] (93%). To observe the effect of review length on the model, different review lengths were applied in respect of the average review length of the dataset category. The accuracy value decreased as the review length increased (95% for less than 10 words, 84% for more than 200 words).

For fake review classification, it was necessary to find a new labelled dataset, the second dataset showed that fake reviews are shorter than original ones with 88% accuracy in fake reviews classification. In the last experiment, an attempt to conduct fake review classification by merging a subset of the first dataset (1001 rows) and a subset of the second dataset (9001 rows) yielded less accuracy with 84% which proves that the supervised learning approach is limited in the face of unlabeled data. The findings can be summarized in these key points:

- The genuine reviews are longer than fake reviews which confirms the initial assumption.
- Positive reviews are longer than negative reviews.
- The Support Vector Machines is a robust model for text classification and opinion spam detection, which confirms the assumption of the second research question.
- The review length affects the Support Vector Machines performance, which confirms the assumption of the first research question.

V. Conclusion

Through the experiments performed in this project, the Support Vector Machines proved to be strong in sentiment analysis and fake review classification. The model's performance was sensitive to the review length due to the number of features considered by the classifier. The lack of real-life datasets impeded the fake review classification in the Amazon reviews dataset, therefore, to make use of the abundance of unlabeled datasets it is needed to employ semi-supervised and unsupervised learning approaches. Although in the literature many studies have claimed to build a publicly available large-scale dataset. The second dataset used in this project constitutes of computer-generated reviews labelled fake and unfiltered Amazon reviews labelled original which might have fake reviews initially. The SVM technique resulted in 88% accuracy, while the merged subsets of the two datasets yielded 84% accuracy.

The goal in the future is to find a dataset that reflects real-life consumer fake reviews. The supervised learning was limiting in terms of classifying unlabeled data, in the future it will be best suited to use semi-supervised or unsupervised learning. This project focused on review-level feature "length." In the future, it would be interesting to combine it with behavioral features to detect fake reviews.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

This work is an accumulated efforts for an ongoing research study of the authors both the main and the co-authors. It is a supervisor/ student research study and the contribution for all authors. Myasar Tabany wrote the paper and enhanced the quality of the results while Meriem Gueffal collected the results and proposed with Myasar Tabany the methodology and the plan for the whole research work. The implementation part was mainly done by Meriem Gueffal. The practical implementation has been revised by Myasar Tabany. Both authors had approved the final version.

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[2] Amazon Press center. (2022). Amazon targets fake review fraudsters on social media. [Online]. Available:

https://press.aboutamazon.com/news-releases/news-release-details/amazon-targets-fake-review-fraudsters-social-media

[3] N. Jindal and B. Liu, "Review spam detection," in *Proc. the 16th International Conference on World Wide Web*., 2007, pp. 1189−1190.

[4] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proc. the 2008 International Conference on Web Search and Data Mining (WSDM-2008)*, 2008, pp. 219−230.

[5] G. Fei, S. Wang, B. Liu, and L. Akoglu, "Detecting changed-hands online review accounts," arXiv preprint, arXiv: 2106.15352, 2021.

[6] M. Ott, C. Cardie, and J. Hancock, "Estimating the Prevalence of deception in online review communities," in *Proc. the International World Wide Web Conference Committee (IW3C2)*, 2012.

[7] The FTC Act. the Office of the Law Revision Counsel. 15 USC code 45. Unfair methods of competition unlawful; prevention by Commission. [Online]. Available: https://www.law.cornell.edu/uscode/text/15/45

[8] The Federal Trade Commission FTC. (2021). FTC puts hundreds of businesses on notice about fake reviews and other misleading endorsements. [Online]. Available: https://www.ftc.gov/news-events/news/press-releases/2021/10/ftc-puts-hundreds-businesses-notice-about-fake-reviews-other-misleading-endorsements

[9] Amazon Help and Customer Service. Community Guidelines. [Online]. Available: https://www.amazon.co.uk/gp/help/customer/display.html?nodeId=GLHXEX85MENUE4XF

[10] J. Seo, S. Kim, and M. Youn, "Current state, problems and promotion of Coupang," *The Journal of Economics, Marketing and Management*, vol. 6, no. 1, pp. 1–8, 2018.

[11] O. Chapelle, P. Haffner, and V. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1055−1064, 1999.

[12] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What yelp fake review filter might be doing?" in *Proc. the International AAAI Conference on Web and Social Media (ICWSM-2013)*, 2013, pp. 409−418.

[13] W. Zhang, C. Bu, T. Yoshida, and S. Zhang, "CoSpa: A co-training approach for spam review identification with support vector machine," *Information*, vol. 7, no. 1, p. 12, 2016.

[14] Y. Ren and D. Ji, "Learning to detect deceptive opinion spam: A survey," *IEEE Access*, vol. 7, pp. 42934–42945, 2019.

[15] Z. Ya, Z. Qingqing, W. Yuhan, and Z. Shuai, "LDA_RAD: A spam review detection method based on topic model and reviewer anomaly degree," *Journal of Physics: Conference Series*, vol. 1550, no. 2, 022008, 2020.

[16] Q. Zhang, J. Wu, P. Zhang, G. Long, and C. Zhang, "Collective hyping detection system for identifying online spam activities," *IEEE Intelligent Systems*, vol. 32, no. 5, pp. 53−63, 2017.

[17] L. Xiang, G. Guo, Q. Li, C. Zhu, J. Chen, and H. Ma, "Spam detection in reviews using LSTM-based multi-entity temporal features," *Intelligent Automation & Soft Computing*, vol. 26, no. 6, pp. 1375−1390, 2021.

[18] N. Hussain, H. Mirza, I. Hussain, F. Iqbal, and I. Memon, "Spam review detection using the linguistic and spammer behavioral methods," *IEEE Access*, vol. 8, pp. 53801−53816, 2020.

[19] J. Rout, A. Dalmia, K. Choo, S. Bakshi, and S. Jena, "Revisiting semi-supervised learning for online deceptive review detection," *IEEE Access*, vol. 5, pp. 1319–1327, 2017.

[20] Z. Zeng, J. Lin, M. Chen, M. Chen, Y. Lan, and J. Liu, "A review structure-based ensemble model for deceptive review spam," *Information*, vol. 10, no. 7, 243, 2019.

[21] W. Zhang, C. Bu, T. Yoshida, and S. Zhang, "CoFea: A novel approach to spam review identification based on entropy and co-training," *Entropy*, vol. 18, no. 12, 429, 2016.

[22] C. Kale, D. Jadhav, and T. Pawar, "Spam review detection using natural language processing techniques," *International Journal of Innovations in Engineering Research and Technology (IJIERT)*, vol. 3, no. 1, pp. 1−6, 2016.

[23] M. Ott, Y. Choi, C. Cardie, and J. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proc. ACL 2011: HLT*, 2011, pp. 309−319.

[24] N. Kangle, R. Kannan, and S. Vispute, "Application of machine learning techniques for fake customer review detection," *Asian Journal for Convergence in Technology (AJCT)*, vol. 7, no. 3, pp. 13−16, 2021.

[25] R. Taqiuddin, F. Bachtiar, and W. Purnomo, "Opinion spam classification on steam review using support vector machine with lexicon-based features," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 6, no. 4, pp. 269−276, 2021.

[26] T. Wang and H. Zhu, "Voting for deceptive opinion spam detection," arXiv preprint, arXiv:1409.4504, 2014.

[27] A. Ligthart, C. Catal, and B. Tekinerdogan, "Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification," *Applied Soft Computing*, vol. 101, 107023, 2021.

[28] M. Patil, S. Nikumbh, A. Parigond, and M. Patil, "Fake product monitoring and removal for genuine product feedback," *International Journal of Emerging Science and Engineering (IJESE)*, vol. 7, no. 1, pp. 1−3, 2021.

[29] Y. Padma and Y. Krishna, "An automatic framework for document spam detection using enhanced context feature matching," *International Journal of Advanced Research in Computer Science*, vol. 9, no. 1, pp. 587−591, 2018.

[30] J. Ye, S. Kumar, and L. Akoglu, "Temporal opinion spam detection by multivariate indicative signals," in *Proc. Tenth International AAAI Conference on Web and Social Media*, 2016.

[31] F. Khurshid, Y. Zhu, Z. Xu, M. Ahmad, and M. Ahmad, "Enactment of ensemble learning for review spam detection on selected features," *International Journal of Computational Intelligence Systems*, vol. 12, no. 1, pp. 387−394, 2019.

[32] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, "Detection of opinion spam based on anomalous rating deviation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8650−8657, 2016.

[33] H. Li, G. Fei, S. Wang, B. Liu, W. Shao, A. Mukherjee, and J. Shao, "Modeling review spam using temporal patterns and co-bursting behaviors," arXiv preprint, arXiv:1611.06625, 2016.

[34] S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, "NetSpam: A network-based spam detection framework for reviews in online social media," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1585−1595, 2017.

[35] N. Long, P. Nghia, and N. Vuong, "Opinion spam recognition method for online reviews using ontological features," *Tạp chÍ Khoa Học*, vol. 61, 44, 2014.

[36] V. Gupta, A. Aggarwal, and T. Chakraborty, "Detecting and characterizing extremist reviewer groups in online product reviews," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 3, pp. 741−750, 2020.

[37] M. Fazzolari, F. Buccafurri, G. Lax, and M. Petrocchi, "Experience: Improving opinion spam detection by cumulative relative frequency distribution," *Journal of Data and Information Quality*, vol. 13, no. 1, pp. 1–16, 2021.

[38] A. Salunkhe, "Attention-based bidirectional LSTM for deceptive opinion spam classification," arXiv preprint, arXiv:2112.14789, 2021.

[39] A. Mukherjee, "Extracting aspect specific sentiment expressions implying negative opinions," in *Proc. International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 194−210, 2016.

[40] H. Li, Z. Chen, A. Mukherjee, B. Liu, and J. Shao, "Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns," in *Proc. ICWSM 2015*, 2015.

[41] J. Salminen, C. Kandpal, A. Kamel, S. Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," *Journal of Retailing and Consumer Services*, vol. 64, 102771, 2022.