Project Report: Detecting Fake Reviews by Feature Engineering and Machine Learning

Chirag Daryani
cdaryani@ualberta.ca
University of Alberta
Edmonton, AB, Canada

Seeratpal Jaura
seeratpa@ualberta.ca
University of Alberta
Edmonton, AB, Canada

ABSTRACT

Reviews capture the opinion of people on a product or service. These reviews help other consumers or businesses to get insights about certain products. However, falsifying these reviews can deceive people into making wrong decisions. Fake review detection continues to remain a challenging task today. This work focuses on detecting fake reviews using an aggregation approach that provides high performance while still attempting to retain explainability.

ACM Reference Format:
Chirag Daryani and Seeratpal Jaura. 2022. Project Report: Detecting Fake Reviews by Feature Engineering and Machine Learning. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Online reviews on different products serve as electronic opinions and they play a significant role in influencing other people to make decisions about the purchase of a product[9] [18]. Ideally, reviews should reflect the honest opinion of real consumers for a product or a service. Unfortunately, we are witnessing a tremendous growth of fake reviews on online platforms [38]. Fake reviews have the potential to negatively impact businesses [1]. They could mislead people into making wrong decisions which would cause monetary loss for both the customers and the product or service providers.[28].

The aim of this project is to detect fake reviews with high performance which continues to be a challenging task. We hypothesize that it is crucial that fake reviews should be detected with high recall (fewer false negatives) because if a fake review gets mislabeled as genuine, it could mislead people or businesses into making certain decisions that can have a significant impact. Not only this, but we also need to provide an intuition for each prediction so that the online review platforms have some substantial justification that would allow them to remove such fake reviews. Hence, we also aim to analyze the relationship between different features, with the goal of better understanding the behavior and patterns of fake reviews.

To achieve this objective, we explore different strategies which include data analysis, rule-based approaches, machine learning, and a combination of both machine learning and rule-based approaches. Our intuition behind performing aggregation is that rule-based methods help to achieve interpretability while machine learning would allow us to generalize the predictions. Therefore, the aggregation mechanism would take into account the decision of both the rule-based approaches and machine learning before making a final prediction.

The contributions of this project are as follows:

- Our project conducted an extensive data analysis to explore what factors could be used to construct simple rule-based methods that are very intuitive. These factors also help us understand the patterns and behaviors of fake reviews and how fake reviews can be distinguished from genuine reviews.
- Our project achieves high detection performance by experimenting with different machine-learning techniques on both the text and non-textual review and reviewer data.
- We then explore methods to combine both the rule-based and machine-learning approaches to get the best of both approaches. Our experiments took into account different ways in which these approaches could be combined and also investigate a custom aggregation approach.

Our results indicate that the combination of both the rule-base and machine learning approach could yield better results as it takes into account the decisions of different classifiers. Furthermore, the combination could help in achieving generalizability while retaining interpretability which is a requirement for this task.

The remaining paper is organized as follows. Section 2 describes the related work, followed by a description of the dataset in Section 3. After that, section 4 of this paper describes the approach in detail which covers the data analysis, rule-based techniques, machine learning approach, and aggregation approaches. Section 5 details the evaluation metrics, the experiments conducted, and their results. In section 6, we provide a discussion and specify some of the future work. Section 7 covers the conclusion. The last section talks about the role of each team member in this project.

2 RELATED WORK

Different approaches have been explored over the past that attempt to solve the problem of detecting fake reviews.

Work done by Mohawesh et al. [20] used the natural language processing technique for fake reviews detection by employing a transformer [30] architecture. Specifically, the authors of this paper attempt to explore attention-based transformer models [2, 15, 34], and an ensemble approach. Their aim was to investigate if the ensemble of transformer models could be a better technique to detect fake reviews. Their results indicate that their method outperformed the state-of-the-art during the time of writing their paper and achieves better accuracy compared to other machine and deep

Conference'17, July 2017, Washington, DC, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

learning approaches. However, these models often refer to as "black box" [27] as it is difficult to interpret the output decision with these models. Therefore, our project focuses on both interpretability and performance.

Previous work done by [10] explored natural language processing techniques on textual features which is similar to some of the methods we explore in our project. Specifically, their project considers emotion, and embedding representation while also employing neural network models.

Another work done by Hassan Sohan et al. [11] explores a machine-learning approach for detecting fake reviews. Their approach is semi-supervised and their results show that Random Forest [4] outperforms with an f-score of 98% [11]. Furthermore, the work done by Villagracia Octaviano [31] also explored feature engineering and machine learning algorithms for fake review detection. As a result, [31] was able to acquire 96.6% accuracy using Extreme Gradient Boosting [5] and also able to extract two new features. These works in machine learning influence us to explore the machine learning approach for fake reviews.

Yu et al. [36] explore graph-based learning methods and detect the relationship between the different attributes of fake reviews. Specifically, their method includes semi-supervised and unsupervised learning approaches. Their work also discusses some open issues, and one among them is explainability which our work attempt to take into account.

In the past, rule-based techniques were also used for this task. Jindal and Liu [12] explore web spam by analyzing the data from Amazon.com. Their work detects spam reviews by detecting duplicate reviews, and doing feature identification by taking into account "review-centric", "reviewer-centric" and "product-centric" features [12]. Their work motivates us to use rule-based approaches to understand user behavior. Although this influenced us in going a certain direction, however, we attempt to design rules based on our own data analysis for the Yelp dataset [25] as described in the next section.

3 DATASET

Our project uses the YelpZip dataset [25], which contains about 608,598 reviews from Yelp.com[1]. These reviews are written for 5,044 restaurants and are written by 260,277 reviewers. It contains features like the text of the review, the date when the review was written, user information, product information, rating given by the user to the product, etc. Based on the anti-fraud filter that Yelp[2] uses, each of the reviews has been labeled as either genuine or fake. We would use these labels as the ground truth for our experiments.

4 APPROACH

This section describes the approach. We divide this section into data analysis, a rule-based approach, machine learning, and aggregation-based techniques.

4.1 Data Analysis

In this section, we would analyze the differences in the characteristics of fake and genuine reviews. We would also utilize the reviewer metadata information in order to analyze if the behavior of fake and genuine reviewers is significantly different or the same as each other. Our aim is to derive useful insights that could be utilized to construct simple rule-based classifiers that can provide a decision about each review.

4.1.1 Length of the Review. First, we wanted to check if the length of the review can be used to distinguish between fake and genuine reviews. For performing this analysis, we split the text of each review into multiple words, and then we counted the number of words in each sentence. This was called the length of the review. On looking at the box plot and the summary statistics of this derived feature, we observed that most of the reviews have a length in the range of 5 to 150 words. We also noticed that some reviews were extremely long with the total number of words exceeding 210. The maximum length of the review was observed to be 612 words.

Our intuition was that fake reviews are more likely to be shorter. The reasoning behind this intuition is that the people who write fake reviews have not actually used the product/service of the restaurant. Due to this, they have extremely limited knowledge of the product and hence they don't have a lot of content or personal experience that they can write about in the review. We also believe that the fake reviewers cannot spend a lot of time reading other reviews and researching the experiences of other people in order to generate content for their fake reviews. Hence, we strongly believed that fake reviews are more likely to be shorter and genuine reviews are more likely to be longer.

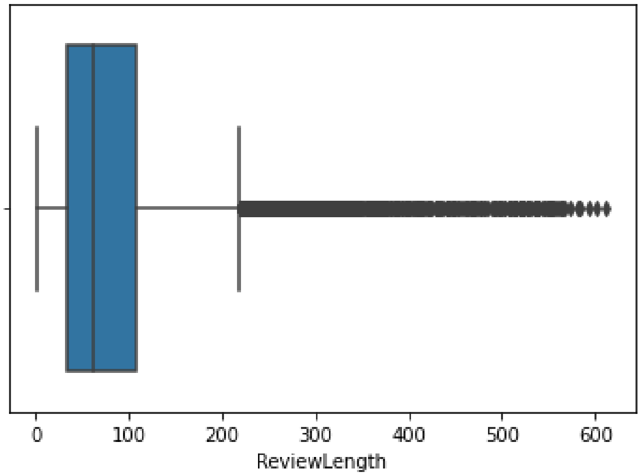

Figure 1: Distribution of Review length

To verify this intuition, we created two groups of reviews. First, is the fake reviews who have been assigned a label of 1 by Yelp's anti-fraud filter. The second group is the genuine reviews who the same filtering algorithm has assigned a label of 0. In order to make the results more interpretable, we decided to create two categories of reviews. If a review has a length of less than 61, then it would be categorized as a "short" review. Otherwise, the review will be considered a "long" review. The number 61 was decided after experimenting with different thresholds. This number is very close

[1] http://odds.cs.stonybrook.edu/yelpzip-dataset/

[2] https://scholar.smu.edu/cgi/viewcontent.cgi?article=1029&context=datasciencereview

to the median length of all the reviews and we believe it is a good number to base our categorization on. After this categorization, we see the percentage of fake and genuine reviews that have been labeled short or long. It was observed that fake reviews tend to have a much more percentage of short reviews (65%) as compared to the percentage of long reviews (35%). On the other hand, the percentage of long reviews (55%) is more in the case of genuine reviews. The results of our analysis verified our intuition regarding the length of the review and we believe it would be a useful feature for building one of the rule-based classifiers.

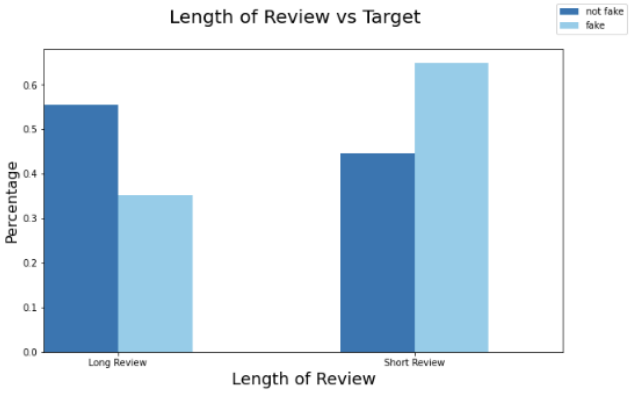

Figure 2: Review length as per Target

4.1.2 *Usefulness of the Review.*

Next, we want to see the impact of the "useful count" which represents the number of users on the platform who have rated the particular review to be helpful. For performing this analysis, we wanted to see the distribution of this numeric feature separately for the group of fake reviews and the group of genuine reviews. To visualize the distribution, we plotted the boxplots of the usefulCount feature. We observed that for fake reviews, the useful count ranges from 0 to 7 with most of the values between 0 and 2. On the other hand, the distribution is significantly different in the case of genuine reviews. For them, the useful count has a much broader range from 0 to 316 votes. We believe this significant difference in the range of values between the two groups would help us make a very good rule-based classifier that can correctly classify the reviews into the two categories based on an appropriate threshold on this feature.

4.1.3 *Rating of Review.*

Each review in the dataset has a rating value that represents how satisfied the reviewer was with the service of the restaurant. There are five possible ratings with 1 being the lowest possible rating and 5 being the highest. We wanted to analyze whether the distribution of these review ratings could be a useful factor to distinguish between the two types of reviews. For performing this analysis, we apply a similar strategy as before whereby we are looking at the percentage of each rating for the fake and genuine reviews.

It was observed that fake reviews tend to have a higher percentage of ratings 1 and 5 as compared to the group of genuine reviews.

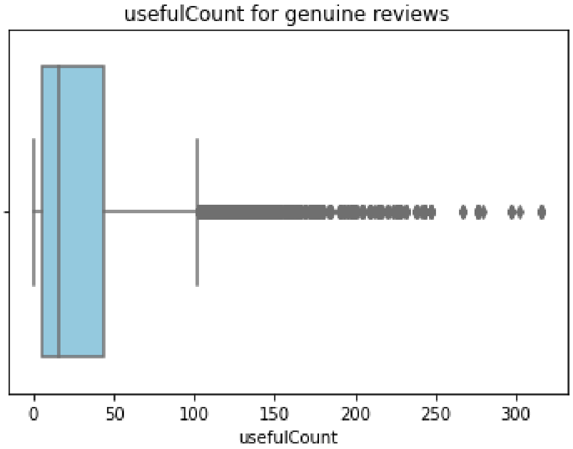

Figure 3: Useful count: Distribution for Genuine Reviews

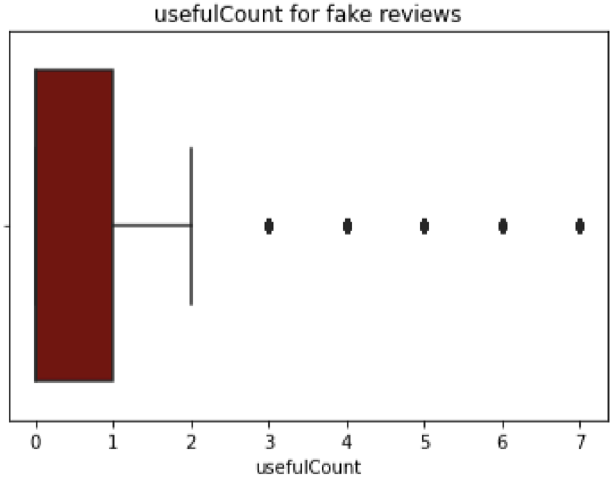

Figure 4: Useful count: Distribution for Fake reviews

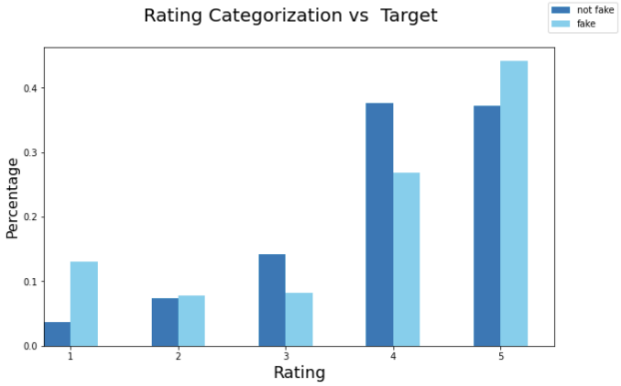

Figure 5: Rating Values (Original) vs Target

For the rating of 2, the percentage is nearly the same in both groups, while fake reviews have a lower percentage of reviews with a rating of 4 (26%) as compared to genuine reviews (37%).

We realized that we need to make the results more interpretable so that we can arrive at a simple rule that uses the review rating for providing a decision. For this, we decided to aggregate the ratings into 3 categories: Low ratings (1), Moderate ratings (2/3/4), and High ratings (5). We wanted to see if this categorization can further distinguish the two groups of reviews.

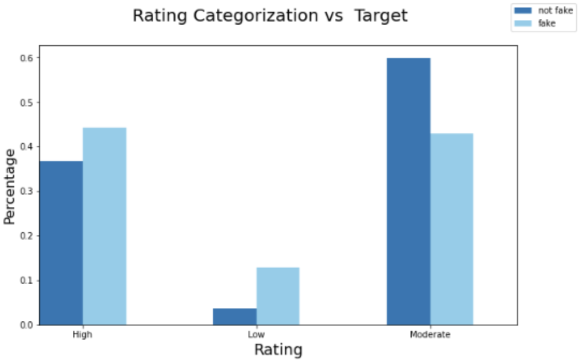

Figure 6: Rating Categories (Derived) vs Target

We observed that fake reviews tend to have a much more percentage of low ratings (13%) as compared to the percentage of low ratings (3%) in genuine reviews. A similar trend is observed in the case of high ratings where fake reviews have more high ratings (44%) as compared to genuine reviews (37%). On the other hand, we observe the opposite trend for moderate ratings. Fake reviews tend to have less percentage of moderate ratings as compared to genuine reviews with the percentages being 42% and 59% respectively.

In our opinion, the justification for this trend can be derived from the human behavior of satisfaction and expectation. Genuine reviewers are more likely to not be extremely satisfied with a product and they always keep a scope for improvement while giving ratings. Hence, they tend to give more moderate ratings. Fake reviewers on the other hand tend to be completely in favor or against a product/service and hence they are more likely to be on the extreme end of the rating spectrum.

4.1.4 Previous Engagement by Reviewer. An important factor that we wanted to analyze was the level of past interactions of a reviewer. Interactions here refer to the number of reviews a particular reviewer has written in the past. We wanted to investigate whether fake and genuine reviews are written by more or less frequent reviewers or if there is no significant difference among the groups.

Our intuition was that fake reviews are less likely to be written by a more frequent reviewer. The reasoning behind this intuition is that more frequent reviewers are authentic customers. They have been on the platform for a long time and have been constantly sharing their experiences with other people. On average, they are more likely to share their true thoughts rather than attempt to attack/promote a particular restaurant. On the other hand, less frequent reviewers don't have any past interaction with the platform and are more likely to have a hidden agenda for/against a restaurant. Hence, genuine reviews are less likely to be written by less frequent reviewers.

To verify this intuition, we created two groups of reviewers based on the reviewCount feature which represents the number of posted reviews by a particular reviewer. If a reviewer has a review count of greater than 30, we consider the reviewer to be a "Trusted" reviewer. If the review count is equal to or less than 30, we consider the reviewer to be "Unverified". We agreed on this number after looking at the distribution of the feature and experimenting with various possible numbers. After this categorization of reviewers, we see the percentage of fake and genuine reviews that have been written by each of these two categories of reviewers.

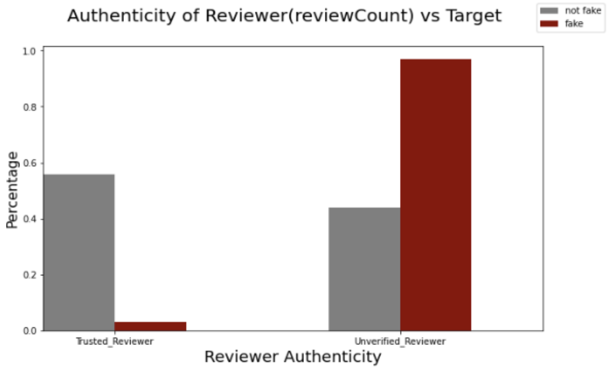

Figure 7: Reviewer Authenticity

We observed that there is a very large difference in percentages for fake reviews. These types of reviews tend to be mostly written by unverified reviewers (97%) as compared to trusted reviewers (3% only). On the other hand, we observe the opposite trend for genuine reviews. These are more likely to be written by trusted reviewers (56%) as opposed to unverified reviewers (44%). These trends give an indication that this aspect of the reviewer metadata can be useful for creating a rule-based classifier.

4.1.5 Popularity of Reviewer. Another important factor is the popularity of the reviewer. We wanted to analyze if the number of friends a reviewer has on the platform influence the distribution of fake and genuine reviews. Our intuition was that fake reviews are less likely to be written by a user who has a lot of friends on the platform. The reasoning behind this intuition is that if a reviewer has a lot of friends, it indicates that they are real people who are known by other users on the platform. The fact that they have a lot of friends demonstrates that a lot of people can vouch for the reliability of that particular user. These users can in fact even be influencers on the platform. Usually, such users would not compromise their reputation by writing fake reviews. Hence, we believed that fake reviews are less likely to be written by reviewers with more friends and more likely to be written by reviewers with few or no friends.

Similar to the previous analysis, in order to verify this intuition, we used the number of friends a reviewer has in order to create these two groups of reviewers. If a reviewer has more than 30 friends, we consider the reviewer to be a “Popular” reviewer. If the number of friends is equal to or less than 30, we consider the reviewer to be “Unknown”. We then saw the percentage of fake and genuine reviews that have been written by each of these two categories of reviewers.

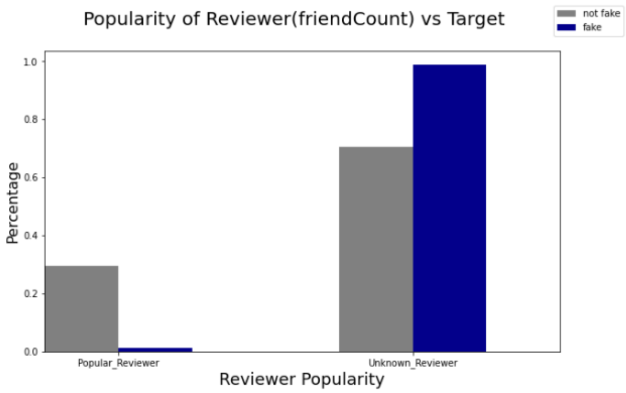

Figure 8: Reviewer Popularity

Exceeding our expectations, we observed that a very large proportion (98.7%) of fake reviews have been written by unknown reviewers with few or no friends. Only 1.3% of the fake reviews were written by popular reviewers with more than 30 friends. For genuine reviews, we were expecting to have more proportion of popular reviewers but surprisingly these reviews are also written by more unknown reviewers (70%) as opposed to popular reviewers (30%). We think that this is the case because the proportion of popular reviewers (23%) is less than the proportion of unknown reviewers (77%). Despite that, the significant difference among the groups in the case of fake reviews indicates that this can be a useful feature for building a rule-based classifier.

4.1.6 Repeat Reviewers for Restaurants. While performing the data analysis, we observed there were some reviewers that had written reviews for multiple restaurants in our dataset. While the review count feature represented the total number of reviews a user had posted across categories on the platform, we wanted to do another analysis in which we restrict the scope, whereby if a reviewer has multiple reviews for restaurants included in our dataset, only then they would be considered repeated reviewers. In order to derive the list of such repeated reviewers, we first grouped the reviews by reviewer id and then calculated the count of the number of reviews for each unique reviewer. After sorting the reviewers according to these counts, we experimented with different thresholds on these counts such that any reviewer with reviews greater than this threshold would be considered a repeated reviewer. We then look at the percentage of fake and genuine reviews among all the reviews that had been written by such reviewers.

We observed that if we only consider the reviews written by reviewers with more than 5 reviews, 99.37% of them would be genuine and only 0.63% of them would be fake. When we started decreasing this threshold by 1 each time, we observed that the percentage of genuine reviews kept decreasing and the percentage of fake reviews kept increasing. For the threshold of 2, we observed that 92% of the reviews were genuine and 8% were fake, which is still a considerable difference.

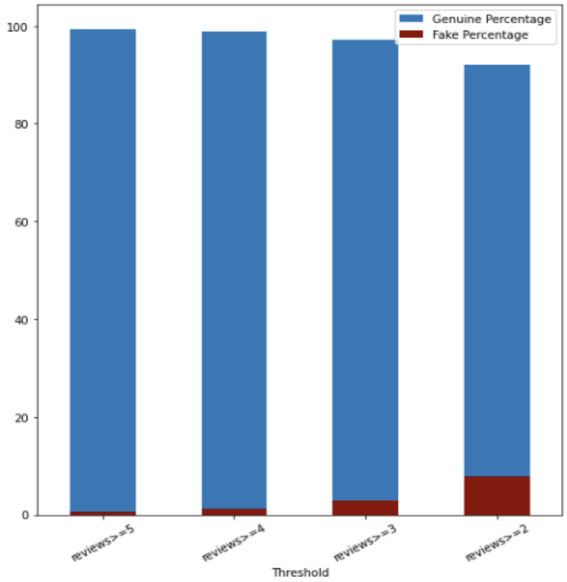

Figure 9: Repeated Reviewers

Apart from the above, we also wanted to analyze whether a repeated reviewer had written multiple reviews for the same restaurant in a short period of time. Our intuition was that such a reviewer would be very likely to be a fake reviewer as no genuine person would write multiple reviews for the same restaurant that they have rated already just a few days ago. Unfortunately, when we tried to derive this feature, we saw that in the dataset, there was no reviewer who had written multiple reviews for the same restaurant. Hence, though this particular analysis is out of scope for this particular dataset, we believe a similar analysis would be useful for other fake review detection tasks.

4.2 Rule based

During our data analysis, we observe some interesting patterns in our data, which influence our decision to investigate the performance of the rule-based approach. In the past, different research works [7, 13, 16] have explored rule-based classifiers; It is to be noted that we derive these rules based on our data analysis. The following rule classifiers were used during our experiments:

- From the data analysis, we observe that extreme rating such as 1 or 5 is mostly flagged as "fake". Therefore, using this pattern the rating-based rule is investigated such that if a review has an extreme rating, it should be classified as fake.
- If a review is shorter than a threshold number of words, then it should be classified as fake. For our experiments, the threshold value is 61 words.

- To analyze if the reviewer is usually active, we observe the number of reviews posted by a reviewer which we refer to as a "repeated reviewer". If a review has been posted by a repeated reviewer but with fewer than 3 reviews, it is likely a fake review.
- Depending on the number of friends, a rule checks if a review has been written by a reviewer with less than 27 friends, then it could be classified as fake.
- Based on the review count of a reviewer, if a review count is less than 30, it is likely a fake review.
- Useful count on a review is an important feature to detect fake reviews. If a review is not rated useful more than three times, then it is likely a fake review.

4.3 Need of Machine learning

Although the rule-based approach is intuitive and helps to understand the behaviors and patterns in the dataset there are a few limitations with the rule-based approach. Firstly, the rule-based approach involves a cold start problem [32]. If we consider an example of a reviewer who just joined the platform and therefore has very few friends, they will be classified as fake by the friend count-based rule classifier even if the user is genuine. Similarly, other rule-based classifiers like the repeated reviewer, useful count, and others would not work just because the user is new to the platform. Therefore, rule-based classifiers do not work in such situations due to the lack of sufficient information about new reviews/reviewers. Secondly, there is a problem of oversimplification or underfitting as the relationship among all features is not covered. These methods imply strict behavior that the user/review needs to follow in order to detect as genuine or fake.

In order to deal with the limitation of rule-based classifiers, our project also considers machine learning approaches so that the hidden interaction between features could be covered.

4.4 Machine Learning

From our data analysis, we observed that there are mainly two types of features in our dataset, one are textual features (for example, review content) and non-textual features (for example, useful count). Based on these features types, we categorized machine learning algorithms as follows:

- Machine learning on non-textual features
- Machine learning on textual features

The machine learning on textual features was performed by first considering the natural language processing techniques to ensure that the model is able to better capture the patterns in the content of fake reviews. Figure 10 shows the overview of the process of conducting machine learning on textual features. As specified that for processing the text, our project has used techniques like Bag of Words (BoW), Term Frequency – Inverse Document Frequency (TF-IDF), and Word Embeddings - Word2Vec.

4.4.1 Bag of Words. To generate a bag of words, our projects remove the stop words from the text and generate n-gram during tokenization [23, 37].

4.4.2 Term Frequency – Inverse Document Frequency (TF-IDF). TF-IDF is a statistical term that finds the importance of words in documents and across all documents; this would help to capture those words that are important in different fake reviews and would help to solidify the probability when performing machine learning classification on top of it. In the past, TF-IDF has been used to get the relevance of words and documents [24].

4.4.3 Word Embeddings - Word2Vec. Word2Vec is a technique that aims to learn vector representation of the text in such a way that similar text would be close to each other in space dimension. The Word2Vec approach takes a text as input and generates word embedding for words in the text [19].

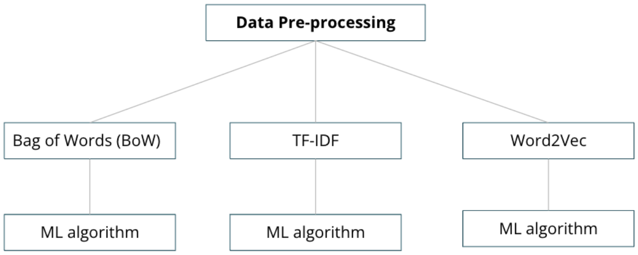

Figure 10: Machine learning on textual features

For our machine learning classification, we have explored different machine learning classifiers such as Support Vector Machine (SVM) [6] [29], Logistic Regression [8], Decision Trees [21], Random Forest [4], Nu-Support Vector Classification [26], XGBoost [5] and Stacking [33] approach.

4.5 Aggregation

To achieve our objective of detecting faking reviews by better understanding the behavior, patterns, and relationship between different features, we explored the aggregation strategy for this task. The aggregation approach aims to combine both rule-based and machine-learning approaches. The reason for choosing the aggregation mechanism is based on the fact the rule-based method gives interpretability while the machine learning models could help to obtain generalizability. Therefore, we explored different aggregation techniques to get the best of both approaches. By combining these approaches through an agreement mechanism, we aim to arrive at a single decision. Some previous works [14], [22] have also explored aggregation mechanisms like majority voting. However, our approach combines rule-based classifiers with machine learning on the non-textual feature. Moreover, we also attempt to explore a custom aggregation approach.

Figure 11 below shows the overview of the aggregation approach.

As indicated that both rule-based and machine learning approach takes input from the dataset, then the approaches are combined based on different mechanism like majority voting and by applying weighing schema on voting to get the final prediction.

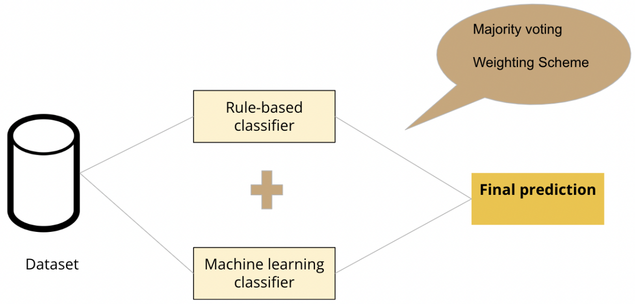

Figure 11: Aggregation Mechanism

Below are the three aggregation strategies explored in this task.

- Majority voting
- Weighing Schema with majority voting
- Custom aggregation

4.5.1 Majority voting. In the majority voting, we incorporate the decision of the rule-based classifier with the best machine learning classifier on non-textual features. As this project investigated the rule based on non-textual features, therefore, the aggregation mechanism also combines with the machine learning model on non-textual features.

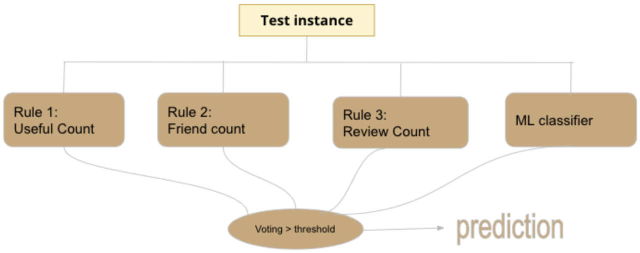

Figure 12: Aggregation Mechanism - majority voting

Figure 12 shows an overview of the majority voting mechanism. In this technique, different rule-based methods are combined with a machine learning model with equal weights. Each approach gives a single category as fake (value 1) and genuine (value 0). We sum these decisions of each classifier, to get the total voting which we then compared against, a threshold value. If the total vote exceeds a certain threshold, we give a prediction as "fake". To illustrate, figure 12 shows the best rule-based approaches with the best ML model. In the past, the work done [35] also explore the majority voting mechanism. However, it is to note that our majority voting schema combines multiple rule base classifiers with a machine learning model.

4.5.2 Weighing Schema with majority voting. This aggregation is leveraging the previous aggregation mechanism by adding weights to each decision classifier. We observe that some classifier tends to perform better as compared to others, therefore, it is crucial to take into account their performance while aggregating these results. Figure 13 shows the overview of weighing schema while conducting voting. In this, each classifier edge has the weight factor (w1,w2,...,wn), these weights decide how much importance should be given to the decision of each classifier.

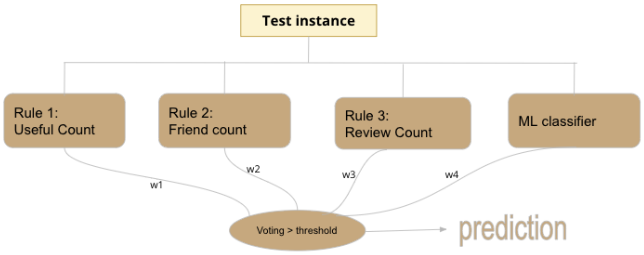

Figure 13: Aggregation Mechanism - Weighing Schema with majority voting

4.5.3 Custom aggregation. The last aggregation that was explored in this project was custom aggregation where the final decision is made based on the decision made at different levels in the tree/chart. Therefore, this aggregation is more like a chart-based aggregation. To illustrate, how this approach works, we combine different rule-based approaches as shown in figure 14. The decision is made at each node. To make the decision we compare the rule-base parameter with a certain threshold value, if the threshold value exceeds, it will predict "fake" right away else the decision falls to another classifier, and the process is repeated until we combine all the best classifiers or we reached the good prediction state.

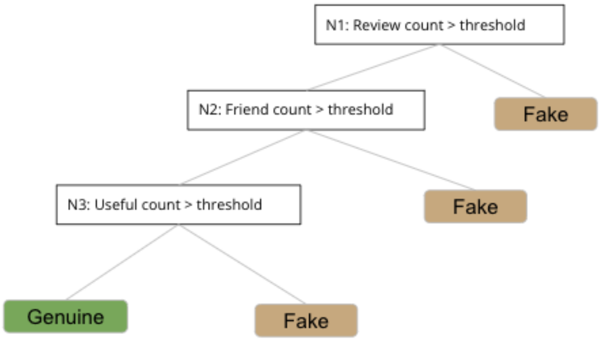
N1: Review count > threshold
N2: Friend count > threshold
Fake
N3: Useful count > threshold
Fake
Genuine
Fake

Figure 14: Aggregation Mechanism - custom approach

Figure 14 shows that the test instance is passed to the first classifier, in this case, a classifier that explores the review count, if the review count is less than a threshold value, then it will detect the test instance as "fake" else it will take another classifier decision into account. The second classifier used is based on the friend count rule, where if a user has a threshold that exceeds the friend count, then "fake" will get predicted otherwise the third classifier will make a decision. The third classifier in this case is a useful count, where the decision is made based on the comparison that if the review has a more useful count than the threshold value, it is likely to be genuine else fake.

We have explored different classifiers by placing them at different levels in the chart and we will discuss in detail those experiments in Section 5.

5 EXPERIMENTS

This section elaborates on the experiments conducted on the above approaches. As part of the data analysis and data preparation for machine learning model training, we performed the standard data preprocessing steps such as cleaning the review text by removing special characters, converting to lowercase, expanding contractions like won't, can't, etc., and removing stopwords using NLTK [17]. We also handle outliers in our numeric features based on the empirical rule. As we expected, the dataset was highly imbalanced with more genuine reviews as compared to fake reviews. To handle this imbalance, we chose the method of undersampling because we had enough data points in each category of the reviews. We partitioned the dataset into 2 subsets, the training dataset with 80% of the records, and the test dataset with 20% of the records. While building our models, we performed extensive hyperparameter tuning for each of the different machine-learning architectures. To be specific, the tuning of hyperparameters of textual and non-textual machine learning approach experiments was considered separately.

5.1 Evaluation metrics

For evaluation purposes, we have considered accuracy, precision, recall, and F1 score as our performance metrics. One of the main focuses of the project is to maximize recall since we believe that it is more important to minimize the false negatives because undetected fake reviews can cause more harm to both the customers and the businesses involved. F1 score is also an important metric since it helps to provide us with a balanced view of performance. Therefore, we consider recall and F1 score as our primary evaluation metrics. All the results we report or these metrics are on the unseen test data.

5.2 Results

We present the results of the rule-based approach in Table 1. It is clearly visible that the performance of different rule-based classifiers varies a lot. Our results indicate that the rule classifier for useful count gives the best accuracy, precision, and F1 score. Although the recall is less compared to some other classifiers like rule-based on friend count, however, it is competitive. Given that all other classifiers except the rule based on the useful count have less accuracy, precision, and F1 score, thus, in our opinion, this rule-based classifier would be preferred for tasks where there is a requirement for balanced performance across the two types of reviews.

For the machine learning approach, the first category of models is machine learning with textual features. For this, we apply natural language processing techniques in combination with multiple machine learning architectures. From the result table 2, it can be viewed that when machine learning is applied on top of the Bag of Words model, then NuSVC (with RBF Kernel) is better compared to other models as it balances out the accuracy, and precision, recall and F1 score. While some other models have high accuracy but low recall or low precision. By analyzing the results of machine learning with the TF-IDF model, it is viewed that the XGBoost has a good performance with competitive accuracy, precision, recall, and F1 score. Although random forest gives the highest recall, however, the accuracy for that model is very low compared to other models. From the results of performing ML models by first applying the Word2Vec model to textual features, it is observed that the Random Forest has the highest recall and F1 score. As one of our objectives is to have few false negatives, therefore, Random forest model which has the highest recall along with the F1 score could be considered a model that works best when machine learning is applied on top of the Word2Vec embeddings.

With respect to the analysis of machine learning models on non-textual features, our results in table 3 indicate that the SVM with a linear kernel function could be beneficial to detect the fake reviews with the highest performance. It is to be noted that the SVM (linear) outperforms other models in terms of accuracy, precision, and F1 score. On the other hand, if we focus only on recall, NuSVC performs the best and would be preferred in situations where the objective is minimizing false negatives. We also explored the feature importance to understand what feature attributes the most while making predictions. From these experiments, we observe that friend count and useful count play a significant role in the final predictions.

Results of aggregation method:

Our project explores three different aggregation strategies: simple majority voting, the weighing schema in majority voting, and custom aggregation. While working with rule-based approaches, we observe that the rules based on the useful count of a review, the number of friends of a reviewer, and a review count give us a good performance. We, therefore, chose these rule-based classifiers in the aggregation. In combination with these rule-based classifiers, we selected the best-performing machine learning classifier as the fourth decision maker.

Table 4 states the results of the aggregation approach using simple majority voting. In this, all the classifiers are assigned equal weights. The decision of all classifiers is put together using a sum which is then compared against the threshold value. Our results indicate that a threshold value of 3 or greater would yield the highest F1 score while balancing recall.

Next, we applied weights to each decision classifier. We experiment with different weights for different classifiers as shown in table 5. Our results indicate that applying more weight to machine learning classifiers than other rule-based classifiers yields good performance. The combination assigning rules 1 and 2 with the weight 1, rule 3 with the weight 2, and the ML classifier with the weight 3, outperforms other combinations in table 5, and in fact, all the combinations that do not apply weights as shown in table 4. This indicates that the weighing-based aggregation strategy could be a potential way to detect fake reviews with high performance.

Our project also explores a custom approach which is a chart-based aggregation as shown in Figure 14. This method provides a very high recall of 99.8% but other performance metrics are lower with the accuracy, precision, and F1 score as 64.4%, 59.0%, and 74.2% respectively. In this case, the threshold for the review count, friend count, and useful count is 30, 27, and 3 respectively.

Rule-based classifiers	Acc	Prec	Recall	F1
If a review has extreme rating, then it should be classified as fake	59.2	60.7	57.4	59.0
If a review is shorter than 61 words, it should be classified as fake	60.2	60.5	64.0	62.2
If a review has been written by a "repeated reviewer" with <3 reviews, it should be classified as "fake"	52.9	52.1	99.8	68.4
If the review has been written by a reviewer with less than 27 friends, it should be classified as fake	65.4	59.8	98.3	74.4
If the reviewCount for the review is less than 30, it should be classified as fake	77.2	70.1	96.8	81.3
If a review has less than 3 votes of being rated as useful, it should be classified as fake	81.8	84.9	78.4	81.5

Table 1: Evaluation results: Rule based approach

Model	ML with BoW				ML with TF-IDF				ML with Word2Vec			
	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1
Random Forest	58.477	55.376	97.246	70.568	58.558	55.460	96.695	70.490	58.783	55.872	92.519	69.670
XGBoost	67.418	66.618	72.856	69.597	66.975	65.003	76.868	70.439	64.867	65.214	67.165	66.175
Stacking	68.505	67.964	72.777	70.288	68.586	69.407	69.079	69.242	65.269	66.113	65.905	66.009
NuSVC with RBF Kernel	66.774	64.748	77.025	70.355	68.344	69.893	67.033	68.433	64.020	66.292	60.393	63.205

Table 2: Evaluation results: ML on textual features

Model	Acc	Prec	Recall	F1
SVM(Linear)	86.0	82.1	92.9	87.2
SVM(RBF)	84.9	79.1	95.7	86.6
NuSVC	82.0	74.7	98.0	84.8

Table 3: Evaluation results: ML on non-textual feature

Combination Parameters	Acc	Prec	Recall	F1
threshold >= 2	78.2	70.9	97.2	82.0
threshold >= 3	84.9	80.5	93.1	86.4
threshold >= 4	82.8	87.8	77.0	82.1

Table 4: Results: Aggregation strategy (Majority voting)

Rule 1	Rule 2	Rule 3	Rule 4 (ML)	Acc	Prec	Recall	F1
3	2	2	3	85.1	80.2	94.1	86.6
1	1	2	3	86.1	82.4	92.5	87.2
2	2	1	4	86.0	82.1	92.9	87.2

Table 5: Results: Aggregation strategy (Weighing schema)

6 DISCUSSION AND LIMITATIONS

Although some of the rule-based approaches achieve high performance compared to the machine learning approaches, the rule-building process very much depends on the data analysis section to figure out the patterns and behaviors. There is certainly scope for designing more rule-based classifiers by performing an even more extensive data analysis. In particular, our project heavily explores non-textual features while having a limited analysis of textual features. This can be an area of future work.

Also, due to time constraints, our project was only able to combine rule-based methods with the machine learning models as part of the aggregation strategy. Our project does not consider more powerful deep-learning-based models and text embeddings that could boost the performance even further when combined with intuitive rule-based classifiers. Therefore, this work could be further leveraged.

7 CONCLUSION

Online reviews influence the decision of customers today and fabricated fake reviews attempt to deceive consumers which often causes financial loss to businesses.[3] Not only this, fake reviews also reduce overall customer satisfaction and hampers the trust of other users of online platforms. Therefore, fake reviews are a rising concern for both consumers and product or service providers.

Our project attempts to solve the task of automatic detection of fake reviews on online platforms. We aim to achieve high performance while still attempting to retain the explainability of simple approaches.

For this purpose, we first perform extensive data analysis and then create different rule-based classifiers. However, rule-based approaches faced some limitations such as underfitting and cold start problems. Therefore, our project also utilizes machine learning algorithms for both textual and non-textual features. Although machine learning approaches help to achieve generalizability, however, they are not as intuitive especially when the number of features is large such as with text data.

Each rule-based classifier and machine learning model has its own strengths and limitations, and thus, in order to get the best of both approaches, we attempt to aggregate both approaches. Our project investigates different techniques like majority voting, weighing schema, and custom-built approach. Our results indicate that weighing based aggregation mechanism helps to achieve the best performance. This also reflects that the right combination of rule-based classifiers and machine learning models would be able to get the best out of both approaches provided proper weights are applied to each classifier.

8 ROLE OF EACH MEMBER

Both team members are working jointly on the project. The below section outlines the task assigned to each member.

Task	Team member
Project Proposal	Chirag + Seeratpal
Proposal Presentation	Chirag + Seeratpal
Literature review	Chirag + Seeratpal
Data analysis	Chirag + Seeratpal
Implementation - Rule-based approach	Chirag + Seeratpal
Implementation - Machine learning approach	Chirag+ Seeratpal
Implementation - Aggregation approach	Chirag+ Seeratpal
Experimentation and tuning	Chirag + Seeratpal
Error analysis	Chirag+Seeratpal
Final presentation	Chirag+Seeratpal
Final project report	Chirag+Seeratpal

REFERENCES

[1] Hadeer Ahmed, Issa Traoré, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy* 1 (2018).

[2] Cesare Alippi, Giacomo Boracchi, and Manuel Roveri. 2011. A just-in-time adaptive classification system based on the intersection of confidence intervals rule. *Neural networks : the official journal of the International Neural Network Society* 24 (06 2011), 791–800. https://doi.org/10.1016/j.neunet.2011.05.012

[3] Supanya Aphiwongsophon and Prabhas Chongstitvatana. 2018. Detecting Fake News with Machine Learning Method. In *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. 528–531. https://doi.org/10.1109/ECTICon.2018.8620051

[4] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[6] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Mach. Learn.* 20, 3 (sep 1995), 273–297. https://doi.org/10.1023/A:1022627411411

[7] Ana Costa, João Guerreiro, Sérgio Moro, and Roberto Henriques. 2019. Unfolding the characteristics of incentivized online reviews. *Journal of Retailing and Consumer Services* 47 (2019), 272–281.

[8] David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* 20, 2 (1958), 215–232.

[9] Paulo Duarte, Susana Costa e Silva, and Margarida Bernardo Ferreira. 2018. How convenient is it? Delivering online shopping convenience to enhance customer satisfaction and encourage e-WOM. *Journal of Retailing and Consumer Services* 44 (2018), 161–169.

[10] Petr Hajek, Aliaksandr Barushka, and Michal Munk. 2020. Fake Consumer Review Detection Using Deep Neural Networks Integrating Word Embeddings and Emotion Mining. *Neural Comput. Appl.* 32, 23 (dec 2020), 17259–17274. https://doi.org/10.1007/s00521-020-04757-2

[11] Md Mahadi Hassan Sohan, Mohammad Monirujjaman Khan, Ipseeta Nanda, and Rajesh Dey. 2022. Fake Product Review Detection Using Machine Learning. In *2022 IEEE World AI IoT Congress (AIIoT)*. 527–532. https://doi.org/10.1109/AIIoT54504.2022.9817271

[12] Nitin Jindal and Bing Liu. 2008. Opinion Spam and Analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (Palo Alto, California, USA) *(WSDM '08)*. Association for Computing Machinery, New York, NY, USA, 219–230. https://doi.org/10.1145/1341531.1341560

[13] Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*. 219–230.

[14] Ludmila I Kuncheva. 2014. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

[15] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H1eA7AEtvS

[16] Huayi Li, Zhiyuan Chen, Arjun Mukherjee, Bing Liu, and Jidong Shao. 2015. Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In *Proceedings of the international AAAI conference on web and social media*, Vol. 9. 634–637.

[17] Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028* (2002).

[18] Sandra MC Loureiro, Luisa Cavallero, and Francisco Javier Miranda. 2018. Fashion brands on retail websites: Customer performance expectancy and e-word-of-mouth. *Journal of Retailing and Consumer Services* 41 (2018), 131–141.

[19] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations*.

[20] Rami Mohawesh, Shuxiang Xu, Matthew Springer, Muna Al-Hawawreh, and Sumbal Maqsood. 2021. Fake or Genuine? Contextualised Text Representation for Fake Review Detection.

[21] Anthony J Myles, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown. 2004. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society* 18, 6 (2004), 275–285.

[22] Shirin Noekhah, Erfan Fouladfar, Naomie Salim, Seyed Hamid Ghorashi, and Ali Akbar Hozhabri. 2014. A novel approach for opinion spam detection in e-commerce. In *Proceedings of the 8th IEEE international conference on E-commerce with focus on E-trust*.

[23] Nidhi A Patel and Rakesh Patel. 2018. A survey on fake review detection using machine learning techniques. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*. IEEE, 1–6.

[24] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. Citeseer, 29–48.

[25] Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*. 985–994.

[26] Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. 2000. New Support Vector Algorithms. *Neural Comput.* 12, 5 (may 2000), 1207–1245. https://doi.org/10.1162/089976600300015565

[27] Yi-han Sheu. 2020. Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research. *Frontiers in Psychiatry* 11 (2020). https://doi.org/10.3389/fpsyt.2020.551299

[28] Edson Tandoc, Zheng Lim, and Rich Ling. 2017. Defining "Fake News": A typology of scholarly definitions. *Digital Journalism* 6 (08 2017), 1–17. https://doi.org/10.1080/21670811.2017.1360143

[29] Vladimir Vapnik. 1999. *The nature of statistical learning theory*. Springer science & business media.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[31] Manolito Villagracia Octaviano. 2021. Fake News Detection Using Machine Learning. In *2021 5th International Conference on E-Society, E-Education and E-Technology* (Taipei, Taiwan) *(ICSET 2021)*. Association for Computing Machinery, New York, NY, USA, 177–180. https://doi.org/10.1145/3485768.3485774

[32] Xuepeng Wang, Kang Liu, and Jun Zhao. 2017. Handling cold-start problem in review spam detection by jointly embedding texts and behaviors. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 366–376.

[33] David H Wolpert. 1992. Stacked generalization. *Neural networks* 5, 2 (1992), 241–259.

[34] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. Curran Associates Inc., Red Hook, NY, USA.

[35] Jianrong Yao, Yuan Zheng, and Hui Jiang. 2021. An ensemble model for fake online review detection based on data resampling, feature pruning, and parameter optimization. *IEEE Access* 9 (2021), 16914–16927.

[36] Shuo Yu, Jing Ren, Shihao Li, Mehdi Naseriparsa, and Feng Xia. 2022. Graph Learning for Fake Review Detection. *Frontiers in Artificial Intelligence* 5 (2022). https://doi.org/10.3389/frai.2022.922589

[37] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics* 1, 1 (2010), 43–52.

[38] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.