AN EMPIRICAL STUDY ON FAKE REVIEW DETECTION
Nhut-Lam Nguyen¹*

Abstract – In recent years, the Internet has opened up opportunities for manufacturers and retailers to advertise and sell their products online. Online shopping is becoming a habit of consumers. Although there are many benefits of buying and selling online such as easy for product selection and comparison from many different sellers before deciding which one to buy. Reading comments before buying a product is a habit of customers. It helps them learn from the experiences of former buyers. However, buying based on product reviews is risky, especially fake reviews. These reviews affect the buyers’ purchase decisions. Detecting fake reviews is a critical problem. In this paper, we proposed a machine learning based framework for detecting fake reviews. We extracted features from text and deployed six machine learning models for classification tasks. Experimental results showed that the SVC is a reliable machine learning algorithm for classifying truthful reviews and fake reviews using the TF-IDF feature extraction technique.

Keywords: fake reviews, fake review detection, opinion spams, term weighting, text representation.

I. INTRODUCTION

With the widespread use of the Internet and social media, people are free and easy to access and share information. In the digital age, e-commerce proves its role in the market economy. E-commerce entails the buying and selling of products or services via the Internet. Vendors have been leveraging the Internet to display and sell their products on their websites or social platforms, such as Twitter and Facebook. Buyers are increasingly purchasing their needed products online. These platforms allow customers to read the comments on the products they intend to buy and comment on the products that they have purchased. These comments are commonly customers’ sharing experiences about products they have purchased. It can be positive, negative, or neutral. The comments influence customer’s purchasing decisions [1]. Roughly 90% of buyers read online reviews of their needed products before making decisions, and 88% of customers trust online reviews [2].

Public opinion has been an important indicator in the e-commerce era. Manipulating online reviews are increasingly influencing customers’ opinions in recent years. As positive or negative comments sway customer purchasing decisions, sellers may ask their employees or even hire people to provide comments to promote their products or attack opponents. These comments are called spam opinions. Spam opinions can be categorized into three groups: the first is fake reviews created to mislead readers; the second type is reviews targeted at the brand instead of the products themselves; the last one is non-reviews such as reviews having no opinion, advertisements, questions, and answers on products [3].

The detection of fake reviews on e-commerce websites has been attracting many researchers. Barbado et al. [4] introduced a fake reviews detection framework in consumer electric retailers. They concluded that Random Forest (RF) can effectively recognize fake reviews with an F-score of 0.82. Elmurungi et al. [5] applied the sentiment analysis technique to detect fake reviews from movie reviews. They used four different classifiers to evaluate the performance of the proposed models, which included Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision Tree

¹Tra Vinh University, Vietnam
*Corresponding author: lamnn@tvu.edu.vn
Received date: 06th July 2023; Revised date: 28th August 2023; Accepted date: 29th August 2023
The SVM model was able to achieve an accuracy of 81.75 without removing stop words. Jia et al. [6] proposed a Latent Dirichlet Allocation linguistic feature based on detecting fake reviews using Yelp Filter Dataset. Their experiment revealed that Logistic Regression (LR) and Multi-layer Perceptron (MLP) can detect fake reviews better compared to SVM. However, Setiawan et al. [7] and Poonguzhali et al. [8] showed that SVM accurately classified fake and non-fake reviews.

Although existing studies have been addressing the fake review detection problem, this paper aims to provide an empirical study on fake review detection using traditional supervised learning algorithms on a well-known dataset. The rest of this paper is organized as follows: section II introduces our fake new detection approach; section III presents our experimental results in detail and section IV concludes the paper.

II. PROPOSED APPROACH

The proposed fake review detection framework is shown in Figure 1. Initially, the dataset for training the classification models was collected. Subsequently, the preprocessing steps, such as the removal of links and stopwords, were performed. For feature extraction, term frequency (TF) and term frequency-inverted document frequency (TF-IDF) methods were employed. The selection of an efficient machine learning model plays a crucial role in the detection framework. This paper evaluated six different machine learning models, including LR, KNN, SVM, RF, Adaboost (ADB), and XG boost (XGB). Based on the experimental results, the model with the best performance was selected as the model for fake review detection.

A. Dataset

In this paper, we used the public dataset collected by Ott et al. [9]. The dataset consists of 800 fake reviews and 800 truthful reviews of 20 hotels in Chicago from TripAdvisor. The fake reviews were collected by Amazon Mechanical Turk. The authors of the dataset only considered the reviews of five stars and a length of at least 150 characters. For each hotel, there were 20 fake reviews and 20 truthful reviews were included. Examples of the reviews in the datasets are shown in Table 1.

Table 1: Examples of reviews

<table>
<thead>
<tr>
<th>Deceptive</th>
<th>Hotel</th>
<th>Polarity</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>truthful</td>
<td>Hyatt</td>
<td>positive</td>
<td>Triple A rate with an upgrade to view rooms was less than $200, including breakfast vouchers. The room offered a great view of the river, lake, and Wrigley Building &amp; Tribune Building. Most major restaurants, shopping, and sightseeing attractions are within walking distance. The room was spacious, and the bed was exceptionally comfortable.</td>
</tr>
<tr>
<td>deceptive</td>
<td>Conrad</td>
<td>positive</td>
<td>We stayed here for a weekend trip to Chicago and we will come back and stay here again. The front desk clerks were helpful. Also, we loved the view. This hotel is a quiet place, up-to-date, and super clean. We highly recommend it and will certainly stay here again.</td>
</tr>
</tbody>
</table>
Nhut-Lam Nguyen

the dataset. The list of stop words can be found here.

C. Text analysis

Word clouds are usually used in text analysis [10]. To explore the text part of the reviews in the dataset, we plot the word cloud of the truthful reviews and fake reviews in Figure 2. It can be seen from the figure that fake and truthful reviews have similar terms. For instance, the terms ‘room’ and ‘hotel’ appear in the reviews frequently. Some words are often used in fake reviews, such as ‘great’, ‘stay’, and ‘service’.

D. Feature extraction

Feature extraction plays an important role in classification models. TF and TF-IDF are two common techniques for feature extraction from text. In this paper, these two techniques were used to extract features from the text field of the reviews in unigram model.

The TF of term \( t \) in review \( r \) is defined as:

\[
TF(t, r) = \frac{n(t, r)}{n(r)}
\]

Where \( n(t, r) \) is the number of times term \( t \) occurs in review \( r \) and \( N(r) \) denotes the total number of terms in review \( r \).

The IDF is calculated as:

\[
IDF(t) = 1 + \log \frac{NR}{NR(t)}
\]

Where \( NR \) is the total number of reviews in the dataset. \( NR(t) \) denotes the number of reviews that term \( t \) appears.

The TF-IDF of a term is calculated as:

\[
TF - IDF = TF \times IDF
\]

In addition to features extracted from text, the polarity of text is also fed to the classification models.

E. Classification models

1) Logistic Regression: LR is a popular supervised machine learning algorithm. LR is used for classifying categorical dependent variables using a set of independent variables. The output of the LR model must be a categorical or discrete value.

2) K-nearest Neighbors: KNN is a nonparametric supervised learning algorithm that can be used for classification and regression problems [11]. The algorithm determines the class label of a test sample by comparing it with the most similar samples in the training set. In this paper, we use the majority voting technique to assign the label for the test sample.

3) Support Vector Machine: SVM is one of the most popular machine learning algorithms recently [12]. The SVM can be used for both classification and regression problems. The algorithm uses the training set to construct hyperplanes that can separate them into different classes. The hyperplane is called the maximum margin hyperplane.

4) Random Forest: RF is a bagging ensemble learning algorithm that uses decision trees as the weak learner to improve classification accuracy.
RF combines the outcomes from multiple decision trees, which are trained using different subsets of the training data, to produce the outcome using majority voting.

5) XG boost: XB is an ensemble machine learning algorithm proposed by Chen and Guestrin [14]. XB is an improvement of gradient boosting in which multiple weak learners are combined into a strong one. An improvement of XB is that the regularization terms are added into the objective function for penalizing the model complexity to avoid overfitting.

6) Adaboost: ADB is a powerful ensemble machine learning algorithm proposed by Freund et al. [15]. It is a boosting algorithm that integrates multiple classifiers, called weak learners, to form a strong classifier.

III. EXPERIMENTS

A. Evaluation metrics

To evaluate the performance of fake review detection models, a confusion matrix was used. In the confusion matrix, true positive (TP) is the total number of reviews that are correctly classified as truthful reviews; false positive (FP) is the fake reviews that are misclassified as truthful ones; true negative (TN) is the fake reviews that are correctly categorized; false negative (FN) denotes the reviews that are incorrectly classified as truthful reviews. The standard metrics used in this paper, derived from the confusion matrix, are defined as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

\[
F1 - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

B. Model performance evaluation

The study trained the models using a five-fold cross-validation method. For fivefold cross-validation, the datasets is divided into five different sets. For each round, four sets are the training set, and the remaining set is the test set. Natural Language Toolkit2, a Python package for natural language processing, was employed to preprocess the text parts. For extracting TF and TF-IDF features, the research used CountVectorizer and TfidfVectorizer from scikit-learn.

The results of using the TF as the review’s feature for different classification models are shown in Table 2. It can be seen from the table that the RF outperforms the rest of the classifiers in terms of all evaluation metrics, followed by the XGB. KNN is the worst model for detecting fake reviews.

While using the TF-IDF as the features, the SVC achieved the best performance compared to other models, as shown in Table 3. The SVC model can classify fake reviews and truthful reviews accurately with an accuracy of 0.879 and a precision of 0.895. LR is also good for detecting fake reviews with an accuracy of 0.876 and a precision of 0.887.

The above analysis indicates that using the TF-IDF feature extraction method, the SVC model is an effective machine learning model for fake review detection, followed by the LR.

Table 2: Model performance with TF

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.747</td>
<td>0.773</td>
<td>0.701</td>
<td>0.734</td>
</tr>
<tr>
<td>KNN</td>
<td>0.649</td>
<td>0.721</td>
<td>0.532</td>
<td>0.583</td>
</tr>
<tr>
<td>SVC</td>
<td>0.731</td>
<td>0.840</td>
<td>0.570</td>
<td>0.676</td>
</tr>
<tr>
<td>RF</td>
<td>0.849</td>
<td>0.860</td>
<td>0.836</td>
<td>0.847</td>
</tr>
<tr>
<td>XGB</td>
<td>0.806</td>
<td>0.810</td>
<td>0.801</td>
<td>0.804</td>
</tr>
<tr>
<td>ADB</td>
<td>0.79</td>
<td>0.794</td>
<td>0.785</td>
<td>0.788</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

This research performed an empirical study on fake review detection issues, using a well-known dataset for training six popular supervised machine learning models. Experimental results
showed that the SVC is a reliable machine-learning algorithm for classifying truthful reviews and fake reviews using the TF-IDF feature extraction technique. In future work, it is advisable to explore other feature extraction from text, such as Bag of Word and Word2Vec, and consider implementing deep learning to address this issue.

**REFERENCES**


---

**Table 3: Model performance with TF-IDF**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.876</td>
<td>0.887</td>
<td>0.852</td>
<td>0.874</td>
</tr>
<tr>
<td>KNN</td>
<td>0.747</td>
<td>0.698</td>
<td>0.876</td>
<td>0.776</td>
</tr>
<tr>
<td>SVC</td>
<td><strong>0.859</strong></td>
<td><strong>0.859</strong></td>
<td>0.878</td>
<td><strong>0.876</strong></td>
</tr>
<tr>
<td>RF</td>
<td>0.831</td>
<td>0.843</td>
<td>0.816</td>
<td>0.838</td>
</tr>
<tr>
<td>XGB</td>
<td>0.818</td>
<td>0.822</td>
<td>0.814</td>
<td>0.816</td>
</tr>
<tr>
<td>ADB</td>
<td>0.790</td>
<td>0.798</td>
<td>0.782</td>
<td>0.788</td>
</tr>
</tbody>
</table>