Survey on Fake Online Review Recognition Techniques
Vaibhavalakshmi. C. D¹, Deepthi. K²
Student¹, Assistant Professor²
Ramaiah Institute of Technology, Bangalore, Karnataka, India

Abstract:
Trust and reputation are important and play a crucial role in empowering several parties to begin a relationships that accomplish mutual advantage, exclusively in an environment of e-commerce (EC). Here one major task is to differentiate among fake reviews and real reviews. This paper gives comparison of best four techniques to identify the fake reviews. The techniques include Naive Bayes, Decision tree, Support Vector Machine (SVM) and K- Nearest neighbor (KNN). Semi supervised learning such as Naive Bayes classifier is used to find spammers. Decision tree algorithm is used to categorize manipulated reviews. SVM is one of the techniques in supervised learning with associated learning algorithms for analyzing dataset utilized for classification. KNN is a learning algorithm and is a non-parametric method for classifying reviews depending on closest training. The review gives outcome as SVM yield the best accuracy for classifying fake and honest online reviews.

Keywords: Spam, online review, Naive Bayes, Decision tree, Bagging and boosting, SVM, KNN.

I. INTRODUCTION

Currently, a huge number of user reviews are taken on almost everything available online in the websites, such as eBay and Amazon etc. Reviews may include product user reviews, intended to assist other users in making their purchasing decisions. There are large number of reviews which make it hard for a customer to go through all and make a decision. Additionally, if the buyer reads some of the product reviews, they find it challenging to discriminate among reasonable and unfair reviews. User reviews are likewise a significant basis of consumer information. However they can intensify or reduce the product or websites reputation based on their credibility.

The reviews are not honestly provided always and fake reviews are present normally. Fake reviews mislead the user to make wrong assumptions about the product. It is hence crucial to classify fake reviews, and is also a stimulating problem in both academia and business. In the previous year researchers have projected numerous fake review discovery methodologies to preserve the accurateness of the results of online opinion mining. The different technologies for online review classification includes Convolution Neural Network (CNN), Big data analytics, Sentiment analysis, K*, Naive Bayes, Decision tree, SVM, KNN, etc. This paper considers the comparison of last four algorithms as they give effective results and more accuracy. The NB classifier is a simple probabilistic classifier depending on the application of the theorem for Bayes.

The NB computes a probabilities set in a given dataset, by combining values. The NB classifier also has fast decision-making process. SVM is a supervised learning model with the associated learning algorithm in machine learning, which analyses data and recognizes patterns used to test classification and regression. A lot of classification algorithms have been proposed recently, but SVM is yet one of the most popular classifiers and is widely used. For all machine learning the K-NN algorithm is a very modest algorithm. The efficiency of the K-NN algorithm is based on several key factors, such as a sufficient distance measure, k parameter and a voting similarity measure. The Decision tree is a machine-learning predictive approach that determines the new sample’s target value on the basis of multiple attribute values of the existing data. This method is suitable for the problem of classification. Percentage split is used as the ideal method in testing option. These four algorithmic techniques are compared with respect to real and fake reviews, accuracy and precision. According to review and comparison SVM gives better performance in all terms.

II. LITERATURE SURVEY

Simran Bajaj et. al., [1] developed a web application which takes users’ inputs (reviews), stores the information in the database and then detects the spammers based on the method proposed. The spam review and a non-review features are discussed in detail and are implemented in practice by constructing own dataset. Here, the reviewer's characteristics, as well as the review content is took into account. The prototype include some features of a reviewer, thus preserving a unique identity for every user. Under some hypotheses it is fairly capable of recognizing spam activities.

Elshrij Ibrahim Elnurngiet. al., [2] proposed a six step approach consisting of supervised sentiment classification methods using Weka tool for classification. The six steps are Collection of Amazon reviews, Cleaning of data, Processing of data incorporating StringToWord vector, Stop words elimination and tokenization and Selection of attribute, Selection of feature, Algorithms for sentiment classification incorporating Naive Bayes, Decision tree, Logistic regression and SVM and the last step is processes for detection.

Chengai Sun et. al., [3] projected a CNN model that captures reviews related to product features by linear product structure and feedback, and then implements a bagging model to bag the CNN model with two documented effectual SVM models to deliver more reliable predictive performance.
The contributions of this work are in particular as follows:

Proposed a new fake review detection model where a CNN model is employed to capture the review features relevant to the product and a classifier is developed on the basis of composition features of the term. Integrate the CNN model with two proficient SVM classification approaches to construct a classification task bagging model to minimize high variance and overfitting of the CNN model.

Harish Baraihyiat et. al., [4] presented an ensemble classification method for fake review identification by grouping the classifiers such as KNN, SVM and Naïve Bayes. The suggested methodology combines the classifiers to resolve the shortcomings of each other, and improves the overall performance.

To estimate the efficiency of the proposed algorithm two different data sets called Ott and Yelp.com are utilized. The algorithm is similarly used in conjunction with various feature extraction procedures to equate its output with the previously proposed procedures. The experimental outcomes demonstrates that the proposed algorithm offers extensive performance improvements over preceding algorithms for balanced and unbalanced class distributions.

Amani Karumanchiet. al., [5] carried out data analysis on the dataset of the Yelp Challenge. The data is analyzed, studied and applied various patterns for spam detection. It was found from the experimental outcomes that, diverse feature sets perform fairly well for derived datasets. The SVM algorithm always works best and provides the best estimation precision.

For example, this study has recognized a threshold of k=5 as a robust suggestion of more precise reviews and this could be utilized in practice. This research definitely leaves a lot of work to be done in the future: comprehensive analysis of review emotions can be useful in ensuring rating accuracy; reviewer travel route should be compared with their history of review; reviewer affinity should be taken into concern when the overall rating of a company is rather than a simple average.

Nidhi A. Patel et. al., [6] conducted a survey on false review detection. The techniques using machine learning for fake review detection are discussed. Analysis of present detection methods and challenges faced in detection are also addressed.

Different types of fake reviews identification methods depending on supervised, semi supervised, as well as unsupervised methodologies are conferred. Different features like verbal features, interactive and social features are seen in detail. Comparison is also done for several techniques to classify fake reviews.

Premadeviet. al., [7] used Naive bayes algorithm to test the proposed feature set and compare it to the neural network algorithm when detecting fraud. The feature set uses the app analysis to decide whether the consumer is committing fraud.

This algorithm helps to identify the feature set with another feature set for fraud detection. Web crawler is used to retrieve feedback from a specific web page. Pre-processing to eliminate stop words is then done. The total division is determined after the categorization of the positive and negative terms using naive bayes approach. The study exhibited that an event like a simple machine learning naive bayes algorithm can display a good result on such an important issue as the identification of fake analysis. Hence, the effects of this research recommend even more, that the machine learning algorithm can be used successfully to address this important issue.

Rajashree Jadhavet. al., [8] examined several techniques for categorizing manipulated reviews and suggests a novel approach to recognize those fake reviews by Decision Tree. This work offers users to recover in general view of the product. Most of them believe that whatsoever the reviews or estimation accessible on internet are honest. But fake information may also be present.

This work discussed an outline of product review mining structure and delivers various characteristics of reviewer and reviews. Additional emphasis is given on the diverse text properties of comments and behavior of the reviewer.

Possible attributes for categorizing reviews are defined. These attributes describe the review. Initially the review content is analyzed, which is associated to the review readability. Next grouping defines attributes associated to sentiment. Attributes related to reviewer and product are also defined.

III. COMPARATIVE ANALYSIS OF FAKE REVIEW DETECTIONTECHNIQUES

The four algorithms are reviewed and comparison is done in terms of the performance. The description and type of the algorithm is shown in table 1. When there is no idea on the dataset, the best algorithm to implement is SVM. SVM is comparatively memory efficient than other approaches considered for fake review detection.

It works well even with semi structured and unstructured data such as trees, images, texts, etc. The risk of over fitting is considerably less in SVM due to generalization of SVM models. When we consider text classification the most appropriate algorithm is SVM. As SVM can handle many features the accuracy of the algorithm is considerably very high. Due to all of the above cons of SVM, SVM is considered to be best among all the four algorithms.

IV. CONCLUSION

This paper depicts different fake review detection methods and comparison is done on four best methods namely Naïve bayes, Decision tree, SVM and KNN. These algorithms can be applied on different datasets.

They can be tested for better performance by using several datasets from online websites such as Ebay, Amazon etc for different products. SVM algorithm gives more accuracy and precision according to survey done.

The confusion matrix is created to identify true positive, true negative, false positive and false negative reviews. From this matrix it will be conveyed that SVM will project the real and fake reviews. In future work, sentiment analysis can be used for classification for detecting manipulated reviews using R tool, Stata, Python and Statistical Analysis System.
Table 1. Manipulated Review Detection Methods

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Explanation</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>Naive Bayes is a simple method for building classifiers and models assigning class labels to problematic cases, defined as feature values vectors, where some finite set drawn the class labels.</td>
<td>Supervised classification</td>
</tr>
<tr>
<td>Decision tree</td>
<td>Decision tree learning is a method to predictive analysis used in analytics, machine learning, and data mining. Using a decision tree, it goes from assumptions on an item to conclusions about the target value of the object.</td>
<td>Supervised classification and clustering</td>
</tr>
<tr>
<td>SVM</td>
<td>Support-vector machines, also support-vector networks are supervised learning representations with related learning algorithms which analyze data utilized for regression and classification analysis.</td>
<td>Supervised classification and regression</td>
</tr>
<tr>
<td>KNN</td>
<td>The KNN algorithm is a non-parametric approach that is used for regression and classification. In both cases, the input contains closest samples of training in the functions pacek. The output is based on whether KNN is used to regression or to classify.</td>
<td>Supervised classification and regression</td>
</tr>
</tbody>
</table>

V. REFERENCES


