Predicting Fake Online Reviews: A Comprehensive Study of Supervised and Semi-Supervised Learning Models

Kondragunta Rama Krishnaiah, Professor, Department of Computer Science and Engineering, R K College of Engineering, Vijayawada - 521456, Andhra Pradesh, India, email: kondraguntark@gmail.com

ABSTRACT

In today's business and commerce landscape, online reviews wield significant influence. Consumers heavily rely on user reviews when making purchasing decisions for products online. Unfortunately, this reliance has led to the rise of opportunistic individuals and groups attempting to manipulate product reviews for their own benefit. To combat this issue, a research paper introduces various text mining models, both semi-supervised and supervised, that aim to detect fake online reviews. The study also compares the effectiveness of these techniques using a dataset known as "Gold Standard." The focus of this research work was on implementing unsupervised machine learning algorithms, such as the expectation maximization-based naive Bayes (EM-NB) and expectation maximization-based support vector machine (EM-SVM). Additionally, supervised machine learning algorithms like NB and SVM models were utilized. To extract features from the dataset, the researchers employed the term frequency-inverse document frequency (TF-IDF) method, which helps uncover relevant properties related to the reviews. The extracted features using TF-IDF were then used to train all the models. After conducting simulations, the results showed that the proposed supervised SVM model outperformed the conventional EM-NB, EM-SVM, and supervised NB models in terms of detecting fake online reviews. This outcome highlights the potential of supervised learning techniques in effectively identifying and addressing fraudulent reviews, thereby bolstering the credibility of online reviews and aiding consumers in making informed decisions.

Keywords: data mining, text mining model, fake online reviews, supervised learning, semi supervised learning.

1. INTRODUCTION

The “phenomenon of fake” is taking over marketing. Major drivers for this are (a) the rapid technological development that enables the creation of artificial consumer-facing outputs, such as deepfakes, and (b) the marketplace evolving around these artificial outputs, related to fake creation, detection, and mitigation. Among the most impactful artificial marketing outputs are fake product reviews — also known as ‘fake reviews,’ ‘deceptive reviews,’ ‘deceptive opinion spam,’ ‘review spam,’ or ‘review fraud’ — that pass as real ones. To this end, studying fake reviews has been suggested as one of the primary agenda items in digital and social media marketing research. Online product reviews, as a form of electronic Word-of-Mouth (eWOM), are major drivers in influencing consumers' purchase decisions. In the United States, more than 80% of consumers indicate they use online reviews before purchasing a product. As reviews are among the most influential factors on consumers' buying behavior, fraudulent actors are tempted to hire writers who specialize in or use automated methods for generating fake reviews to enhance the attractiveness of their products and services, or to degrade competitors' reputation. Fake reviews can be created in two main ways. First, in a (a) human-generated way by paying human content creators to write authentic-appearing but not real reviews of products — in this case, the review author never saw said products but still writes about them. Second, in a (b) computer-generated way by using text-generation algorithms to automate the fake review creation. Traditionally, human-generated fake reviews have been traded like commodities in a “market of fakes”– one can simply order reviews online in a given quantity, and human writers would carry out the work. However, the technological progress in text generation – natural language processing (NLP) and machine learning (ML) to be more specific – has incentivized the automation of fake reviews, as with generative language models, fake reviews could be generated at scale and a fraction of the cost compared to human-generated fake reviews.

This issue is important for marketing and e-commerce domains for three main reasons. First, (a) fake reviews may erode consumer trust in online reviews as a whole, which would signify a major market decline. Sincere consumers write reviews to share their experiences, either positive or negative. Hence, truthful reviewing renders a valuable service in the marketplace, as the information in these reviews provides a signal of quality for other consumers. A truthful marketplace for reviews is also in the interest of companies, as they can receive authentic feedback from customers that can be analyzed to improve products and services. If fake reviews were to permeate the marketplace at scale, this would risk systematically degrading source credibility of online reviews in general. The consequence might be adverse selection, a process in which consumers are unable to distinguish good reviews from bad ones.

2. LITERATURE SURVEY

J. K. Rout et.al proposed with more consumers using online opinion reviews to inform their service decision making, opinion reviews have an economic impact on the bottom line of businesses. Unsurprisingly, opportunistic individuals or groups have attempted to abuse or manipulate online opinion reviews (e.g., spam reviews) to make profits and so on, and that detecting deceptive and fake opinion reviews is a topic of ongoing research interest. In this paper, we explain how semi-supervised learning methods can be used to detect spam reviews, prior to demonstrating its utility using a data set of hotel reviews.

E. P. Lim et.al aimed to detect users generating spam reviews or review spammers. We identify several characteristic behaviors of review spammers and model these behaviors so as to detect the spammers. In particular, we seek to model the following behaviors. First, spammers may target specific products or product groups in order to maximize their impact. Second, they tend to deviate from the other reviewers in their ratings of products. We propose scoring methods to measure the degree of spam for each reviewer and apply them on an Amazon review dataset. We then select a subset of highly suspicious reviewers for further scrutiny by our user evaluators with the help of a web-based spammer evaluation software specially developed for user evaluation experiments. Our results show that our proposed ranking and supervised methods are effective in discovering spammers and outperform other baseline method based on helpfulness votes alone. We finally show that the detected spammers have more significant impact on ratings compared with the unhelpful reviewers.

J. Li, M. Ott et.al focused on consumers' purchase decisions are increasingly influenced by user-generated online reviews. Accordingly, there has been growing concern about the potential for posting deceptive opinion spam - fictitious reviews that have been deliberately written to sound authentic, to deceive the reader. In this paper, we explore generalized approaches for identifying online deceptive opinion spam based on a new gold standard dataset, which is comprised of data from three different domains (i.e., Hotel, Restaurant, Doctor), each of which contains three types of reviews, i.e. customer generated truthful reviews, Turker generated deceptive reviews and employee (domain-expert) generated deceptive reviews. Our approach tries to capture the general difference of language usage between deceptive and truthful reviews, which we hope will help customers when making purchase decisions and review portal operators, such as TripAdvisor or Yelp, investigate possible fraudulent activity on their sites.

M. Ott, Y. Choi et.al focused on consumers increasingly rate, review and research products online. Consequently, websites containing consumer reviews are becoming targets of opinion spam. While recent work has focused primarily on manually identifiable instances of opinion spam, in this work we study deceptive opinion spam---fictitious opinions that have been deliberately written to sound authentic. Integrating work from psychology and computational linguistics, we develop and compare three approaches to detecting deceptive opinion spam, and ultimately develop a classifier that is nearly 90% accurate on our gold-standard opinion spam dataset. Based on feature analysis of our learned models, we additionally make several theoretical contributions, including revealing a relationship between deceptive opinions and imaginative writing.

A. Heydari et.al focused on online reviews have become the most important resource of customers' opinions. These reviews are used increasingly by individuals and organizations to make purchase and

business decisions. Unfortunately, driven by the desire for profit or publicity, fraudsters have produced deceptive (spam) reviews. The fraudsters' activities mislead potential customers and organizations reshaping their businesses and prevent opinion-mining techniques from reaching accurate conclusions. The present research focuses on systematically analyzing and categorizing models that detect review spam. Next, the study proceeds to assess them in terms of accuracy and results. We find that studies can be categorized into three groups that focus on methods to detect spam reviews, individual spammers and group spam. Different detection techniques have different strengths and weaknesses and thus favor different detection contexts.

Research gap

As fake reviews pose a pervasive and damaging problem, helping consumers and businesses differentiate truthful reviews from fake ones remains a vital but challenging task. Fake review detection can combine manual efforts, supervised ML, and heuristic methods. Some approaches in the literature focus solely on features extracted from the review text. Linguistic characteristics range from counting the frequency of words or n-grams to more advanced approaches relying on distributional semantics. However, despite the progress made in detection studies, considerable challenges lie ahead. Classification performance needs improvement to keep up with text-generation algorithms. Datasets may not be appropriately devised, contain mislabeled instances, or are not made publicly available. The key takeaway from previous studies is that automatic fake review detection has been only partially successful. While one study cannot tackle all gaps, our study leverages state-of-the-art NLP technologies to generate a robust dataset for fake review detection and then compare manual (crowdsourcing) and automated (ML algorithm) performance to detect computer-generated fake reviews. We make our experiments available for future development.

Existing System

Researchers have been studying many approaches for detection of these fake online reviews. Some approaches are reviewing content based and some are based on behavior of the user who is posting reviews. Content based study focuses on what is written on the review that is the text of the review where user behavior-based method focuses on country, ip-address, number of posts of the reviewer etc. Most of the proposed approaches are supervised classification models. Few researchers also have worked with semi-supervised models. Semi-supervised methods are being introduced for lack of reliable labeling of the reviews.

Disadvantages

- In the existing work, the system uses only semi-supervised learning.
- Only Text Classification as sentiment text and it never finds fake review.

3. PROPOSED SYSTEM

We have implemented both semi-supervised and supervised classifications. For semi-supervised classification of the data set, we have used Expectation-Maximization (EM) algorithm. The Expectation Maximization algorithm is designed to label unlabeled data to be used for training. The algorithm operates as follows: A classifier is first derived from the labeled dataset as shown in Figure 1. This classifier is then used to label the unlabeled dataset. Let this predicted set of labels be PU. Now, another classifier is derived from the combined sets of both labeled and unlabeled datasets and is used to classify the unlabeled dataset again. This process is repeated until the set PU stabilizes. After a stable PU set is produced, we trained the classification algorithm with the combined training set of both labeled and unlabeled datasets and deployed it for predicting test dataset. The algorithm is given below. As classifier, we have used Support Vector machines (SVM) and Naive Bayes (NB) classifier with EM algorithm. Scikit Learn package of Python programming language provides sophisticated library of these classifiers. Hence for our research work, we have used Python with scikit-learn and numpy packages. We have tuned the parameters of the SVM for better results. For supervised classification, we have used Naive Bayes and SVM classifiers. We know, Naive Bayes classifier can be implemented where conditional independence property is maintained. As text comes

randomly from the user mind, we can't know what the next line and word is going to be. Hence, Naive Bayes classifier is popularly used in text mining. It is a probabilistic method hence it can be used both for classification and regression. It is also very fast to calculate.

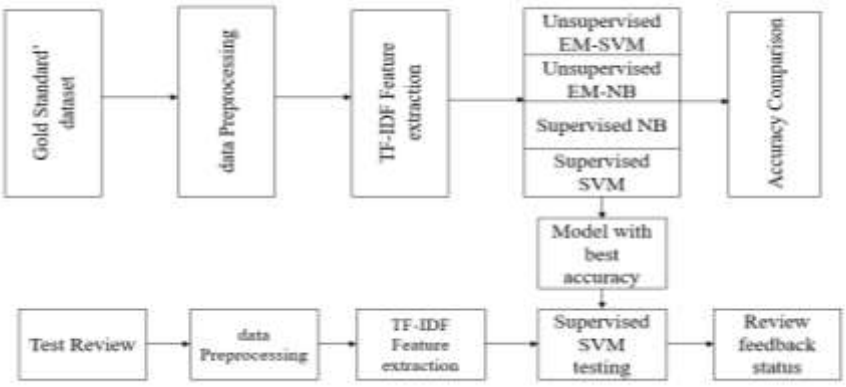

Figure 1. Proposed block diagram

Advantages

- The system is very fast and effective due to semi-supervised and supervised learning.
- Focused on the content of the review-based approaches. As a feature we have used word frequency count, sentiment polarity and length of review.

3.1 Dataset

In this section, we present an overview of the datasets used in this work. These datasets consist of eleven gold standard datasets of short messages, which were labeled by humans as positive or negative according to their sentiment polarity. These datasets consist of data in 9 different languages, besides two sets of messages in English. Their content are from different contexts from Twitter and Website reviews. Smaller datasets contain dozens of instances and some of them few thousands of posts. Random tweets include data of different subjects and the review datasets consist of labeled messages from costumers' reviews about different products and movies. To allow a fair comparison, we selected messages in English from these two groups of data, Random tweets and Reviews. Further, the dataset contains total 1600 reviews and then application using 1280 reviews for training and 320 reviews for testing.

3.2 Preprocessing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

3.3 Splitting the Dataset

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model. Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance.

So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

3.4 TF-IDF Feature extraction

TF-IDF which stands for Term Frequency – Inverse Document Frequency. It is one of the most important techniques used for information retrieval to represent how important a specific word or phrase is to a given document. Let's take an example, we have a string or Bag of Words (BOW) and we have to extract information from it, then we can use this approach.

The TF-IDF value increases in proportion to the number of times a word appears in the document but is often offset by the frequency of the word in the corpus, which helps to adjust with respect to the fact that some words appear more frequently in general. TF-IDF use two statistical methods, first is Term Frequency and the other is Inverse Document Frequency. Term frequency refers to the total number of times a given term t appears in the document doc against (per) the total number of all words in the document and the inverse document frequency measure of how much information the word provides. It measures the weight of a given word in the entire document. IDF show how common or rare a given word is across all documents. TF-IDF can be computed as TF-IDF.

TF-IDF do not convert directly raw data into useful features. Firstly, it converts raw strings or dataset into vectors and each word has its own vector. Then we'll use a particular technique for retrieving the feature like Cosine Similarity which works on vectors, etc.

3.5 Proposed SVM

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

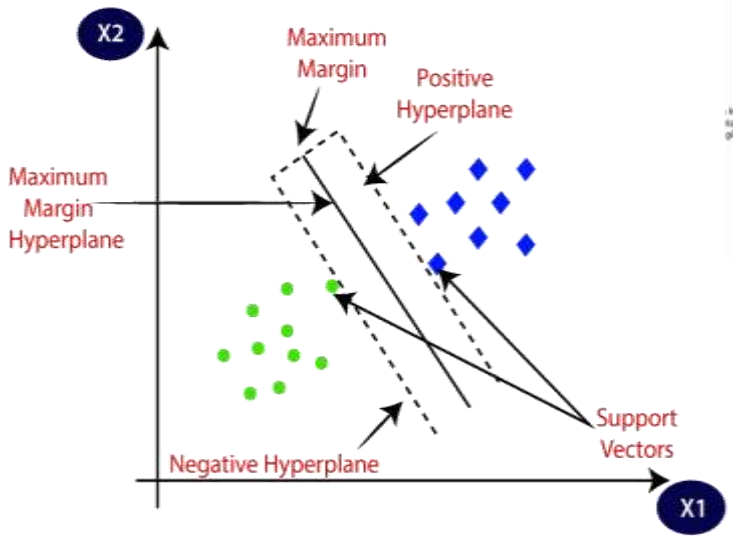

Fig.2. Analysis of SVM

3.5.1 Types of SVM: SVM can be of two types:

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can

be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier

4. RESULTS

This work considered gold standard dataset which contains 1600 reviews from which 800 are genuine reviews and 800 are fake reviews to train both supervised and semi supervised learning approaches. We have two columns in given dataset such as Review and Label where Review column contains user review and label column contains values as 0 or 1 where 0 means FAKE review and 1 means genuine review. After training with the proposed algorithm, we can apply test review on trained model to predict it class as FAKE or GENUINE. Figure 3 shows the accuracy comparison graph. The simulations revealed that the proposed supervised SVM resulted in superior performance as compared to conventional EM-NB, EM-SVM and supervised NB. Figure 4 shows the sentiment graph.

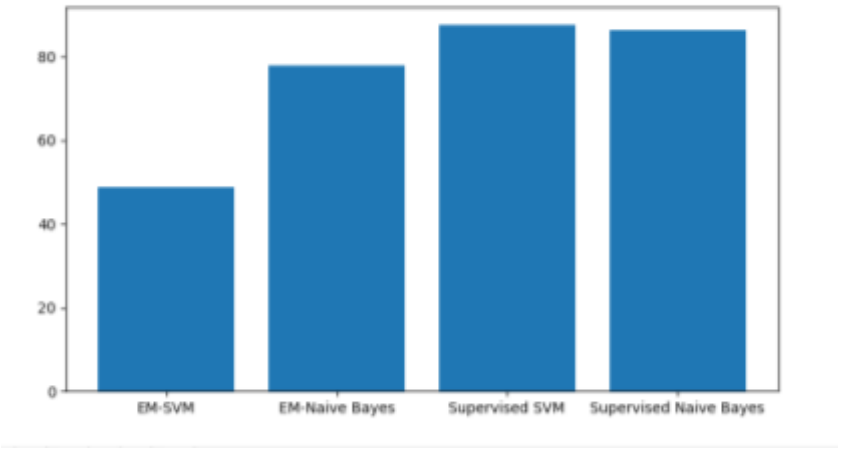

Fig.3. Accuracy comparison graph.

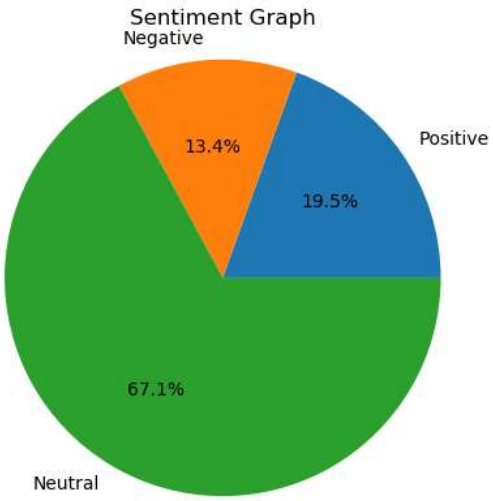

Fig.4. sentiment graph

5. CONCLUSION

We have shown several semi-supervised and supervised text mining techniques for detecting fake online reviews in this research. We have combined features from several research works to create a better feature set. Also, we have tried some other classifier that were not used on the previous work. Thus, we have been able to increase the accuracy of previous semi supervised techniques. We have also found out that supervised SVM classifier gives the highest accuracy. This ensures that our dataset is labeled well as we know semi-supervised model works well when reliable labeling is not available. In our research work we have worked on just user reviews. In future, user behaviors can be combined with texts to construct a better model for classification. Advanced preprocessing tools for tokenization can be used to make the dataset more precise. Evaluation of the effectiveness of the proposed methodology can be done for a larger data set.

REFERENCES

[1] K. D. Lee, K. Han, S. -H. Myaeng, "Capturing word choice patterns with LDA for fake review detection in sentiment analysis," Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics (WIMS '16), 1–7, Association for Computing Machinery, New York, NY, USA (2016)

[2] H. Ahmed, I. Traore, S. Saad, "Detecting opinion spams and fake news using text classification," Security and Privacy, 1 (2018), p. 1, 2018

[3] L. Floridi, M. Chiriatti, "GPT-3: its nature, scope, limits, and consequences," Minds Mach. (2020), pp. 1-14, 2020

[4] M. Ott, Y. Choi, C. Cardie, H. T. Jeffrey, "Finding deceptive opinion spam by any stretch of the imagination," (2011), *arXiv preprint arXiv:1107.4557* (2011)

[5] M. Petrescu, K. O'Leary, D. Goldring, S. B. Mrad, "Incentivized reviews: promising the moon for a few stars," J. Retailing Consum. Serv., 41 (2018), pp. 288-295, 2018

[6] Costa, J. Guerreiro, S. Moro, R. Henriques, "Unfolding the characteristics of incentivized online reviews," J. Retailing Consum. Serv., 47 (2019), pp. 272-281, 2019

[7] S. Kaabachi, S. B. Mrad, M. Petrescu, Consumer initial trust toward internet-only banks in France, Int. J. Bank Market., 35 (6) (2017), pp. 903-924, 10.1108/IJBM-09-2016-0140, January 2017

[8] N. Jindal, B. Liu, "Opinion spam and analysis," Proceedings of the 2008 International Conference on Web Search and Data Mining (2008), pp. 219-230

[9] R. Filieri, "What makes an online consumer review trustworthy?" Ann. Tourism Res., 58 (May 2016) (2016), pp. 46-64, 10.1016/j.annals.2015.12.019.

[10] V. Sandulescu, M. Ester, "Detecting singleton review spammers using semantic similarity," Proceedings of the 24th International Conference on World Wide Web (2015), pp. 971-976

[11] C. Mattson, R. L. Bushardt, A. R. Artino Jr., "When a Measure Becomes a Target, it Ceases to Be a Good Measure," (2021)

[12] D. Plotkina, A. Munzel, J. Pallud, "Illusions of truth—experimental insights into human and algorithmic detections of fake online reviews," J. Bus. Res., 109 (March 2020) (2020), pp. 511-523, 10.1016/j.jbusres.2018.12.009

[13] H. Sun, A. Morales, X. Yan, "Synthetic review spamming and defense," Proceedings of the 22nd International Conference on World Wide Web Companion, 9 (2013), Rio de Janeiro, Brazil

[14] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, H. A. Najada, "Survey of review spam detection using machine learning techniques," Journal of Big Data, 2 (1) (2015), p. 23, 10.1186/s40537-015-0029-9, October 2015

[15] A. Munzel, "Assisting consumers in detecting fake reviews: the role of identity information disclosure and consensus," J. Retailing Consum. Serv., 32 (2016), pp. 96-108, 2016

[16] J. Karimpour, A. A. Noroozi, and S. Alizadeh, "Web spam detection by learning from small labeled samples," International Journal of Computer Applications, vol. 50, no. 21, pp. 1–5, July 2012.