IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 2, October 2023

Spam Review Detection Using Machine Learning

Kiran Naik[1], **Kajal Naik**[2], **Dipti Kapadi**[3], **Devyani More**[4], **Prof. Poonam Dholi**[5]

Students, Department of Computer Engineering[1,2,3,4]

Faculty, Department of Computer Engineering[5]

Matoshri College of Engineering and Research Center, Eklahare, Nashik, Maharashtra, India

Abstract: *With the continuous evolve of E-commerce systems, online reviews are mainly considered as a crucial factor for building and maintaining a good reputation. Moreover, they have an effective role in the decision making process for end users. Usually, a positive review for a target object attracts more customers and lead to high increase in sales. Nowadays, deceptive or fake reviews are deliberately written to build virtual reputation and attracting potential customers. Thus, identifying fake reviews is a vivid and ongoing research area. Identifying fake reviews depends not only on the key features of the reviews but also on the behaviors of the reviewers. This paper proposes a machine learning approach to identify fake reviews. In addition to the features extraction process of the reviews, this paper applies several features engineering to extract various behaviors of the reviewers. The paper compares the performance of several experiments done on a real Yelp dataset of restaurants reviews with and without features extracted from users behaviors. In both cases, we compare the performance of several classifiers; KNN, Naive Bayes (NB), SVM, Logistic Regression and Random forest. Also, different language models of n-gram in particular bi-gram and tri-gram are taken into considerations during the evaluations. The results reveal that KNN(K=7) outperforms the rest of classifiers in terms of f-score achieving best f-score 82.40%. The results show that the f-score has increased by 3.80% when taking the extracted reviewers behavioral features into consideration.*

Keywords: Fake reviews detection; data mining; supervised machine learning; feature engineering

I. INTRODUCTION

Nowadays, when customers want to draw a decision about services or products, reviews become the main source of their information. For example, when customers take the initiation to book a hotel, they read the reviews on the opinions of other customers on the hotel services. Depending on the feedback of the reviews, they decide to book room or not. If they came to a positive feedback from the reviews, they probably proceed to book the room. Thus, historical reviews became very credible sources of information to most people in several online services. Since, reviews are considered forms of sharing authentic feedback about positive or negative services, any attempt to manipulate those reviews by writing misleading or inauthentic content is considered as deceptive action and such reviews are labeled as fake [1]. Such case leads us to think what if not all the written reviews are honest or credible. What if some of these reviews are fake. Thus, detecting fake review has become and still in the state of active and required research area [2]. Machine learning techniques can provide a big contribution to detect fake reviews of web contents. Generally, web mining techniques [3] find and extract useful information using several machine learning algorithms. One of the web mining tasks is content mining. A traditional example of content mining is opinion mining [4] which is concerned of finding the sentiment of text (positive or negative) by machine learning where a classifier is trained to analyze the features of the reviews together with the sentiments. Usually, fake reviews detection depends not only on the category of reviews but also on certain features that are not directly connected to the content. Building features of reviews normally involves text and natural language processing NLP. However, fake reviews may require building other features linked to the reviewer himself like for example review time/date or his writing styles. Thus the successful fake reviews detection lies on the construction of meaningful features extraction of the reviewers. To this end, this paper applies several machine learning classifiers to identify fake reviews based on the content of the reviews as well as several extracted features from the reviewers. We apply the classifiers on real corpus of reviews taken from Yelp [5]. Besides the normal natural language processing on the corpus to extract and feed the features of the reviews to the

DOI: 10.48175/IJARSCT-13504

ISSN 2581-9429 IJARSCT

classifiers, the paper also applies several features engineering on the corpus to extract various behaviors of the reviewers. The paper compares the impact of extracted features of the reviewers if they are taken into consideration within the classifiers. The papers compares the results in the absence and the presence of the extracted features in two different language models namely TF-IDF with bi-grams and TF-IDF with tri-grams. The results indicates that the engineered features increase the performance of fake reviews detection process. The rest of this paper is organized as follows: Section II Summarizes the state of art in detecting fake reviews. Section III introduces a background about the machine learning techniques. Section IV presents the details of the proposed approach..

II. LITERATURE REVIEW

Literature survey is gathering the information of previous work done related to your project. It contains the research study year, researchers name, technologies used and drawback of the system.Detection of opinion spam was first introduced by Jindal & Liu in 2008. They categorized the review spam into 3 categories: Untruthful opinions (if fraudsters write positive fake opinions to promote some targets is called as hyper spam and if fraudsters write negative fake opinions to damage the reputation of some targets is called as defaming spam), reviews on brands only (fraudsters write only about the brand, i.e. the manufacturers of the products rather than the products) and non-reviews (fraudsters write something that is totally unrelated to the products, this may be either advertisements or irrelevant opinion). Authors introduced three types of feature in their proposed work i.e., review centric features, reviewer centric features and product centric features. Lim et al. proposed a model that is based on behavior of spammers. They used to assign a
rank to spammer on the basis of behavior scoring method and they detect spammers according to that rank. Authors collected data set from amazon.com and applied the concept of both behavior scoring method and supervised learning technique to detect review spammers.

III. PROPOSED SYSTEM

3.1 System Model

In above architecture a dataset has been collected and a classification model has been developed and evaluated. The purpose of preprocessing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model.The technique includes data formatting, cleaning, and sampling. A dataset used for machine learning should be partitioned into three subsets training, test, and validation sets. Training set

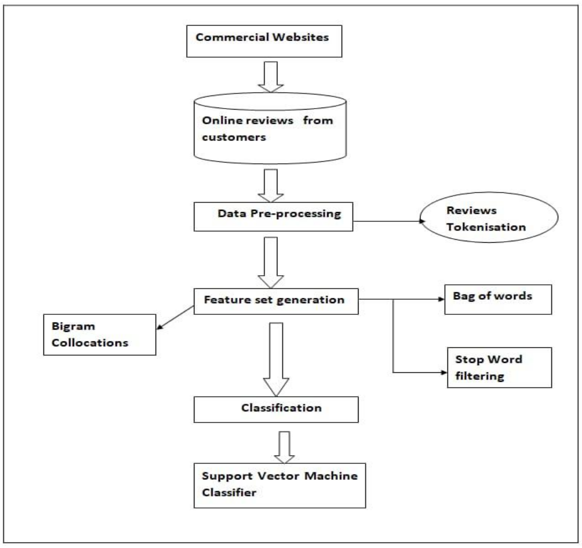

Fig 2 System Architecture for proposed system

DOI: 10.48175/IJARSCT-13504

IV. PROPOSED METHODOLOGY

This section explains the details of the proposed approach shown in figure 1. The proposed approach consists of three basic phases in order to get the best model that will be used for fake reviews detection. These phases are explained in the following:

V. METHODOLOGY

5.1 Data Preprocessing T

he first step in the proposed approach is data preprocessing one of the essential steps in machine learning approaches. Data preprocessing is a critical activity as the world data is never appropriate to be used. A sequence of preprocessing steps have been used in this work to prepare the raw data of the Yelp dataset for computational activities. This can be summarized as follows: 1) Tokenization: Tokenization is one of the most common natural language processing techniques. It is a basic step before applying any other preprocessing techniques. The text is divided into individual words called tokens. For example, if we have a sentence ("wearing helmets is a must for pedal cyclists"), tokenization will divide it into the following tokens ("wearing" , "helmets" , "is" , "a", "must", "for" , "pedal" , "cyclists"). 2) Stop Words Cleaning: Stop words are the words which are used the most yet they hold no value. Common examples of the stop words are (an, a, the, this). In this paper, all data are cleaned from stop words before going forward in the fake reviews detection process. 3) Lemmatization: Lemmatization method is used to convert the plural format to a singular one. It is aiming to remove inflectional endings only and to return the base or dictionary form of the word. For example: converting the word ("plays") to ("play")

5.2 Feature Extraction

Feature extraction is a step which aims to increase the performance either for a pattern recognition or machine learning system. Feature extraction represents a reduction phase of the data to its important features which yields in feeding machine and deep learning models with more valuable data. It is mainly a procedure of removing the unneeded attributes from data that may actually reduce the accuracy of the model [30]. Several approaches have been developed in the literature to extract features for fake reviews detection. Textual features is one popular approach . It contains sentiment classification which depends on getting the percent of positive and negative words in the review; e.g. "good", "weak". Also, the Cosine similarity is considered. The Cosine similarity is the cosine of the angle between two n-dimensional vectors in an n-dimensional space and the dot product of the two vectors divided by the product of the two vectors' lengths (ormagnitudes). TF-IDF is another textual feature method that gets the frequency of both true and false (TF) and the inverse document (IDF). Each word has a respective TF and IDF score and the product of the TF and IDF scores of a term is called the TF-IDF weight of that term . A confusion matrix is used to classify the reviews into four results; True Negative (TN): Real events are classified as real events, True Positive (TP): Fake events are classified as fake, False Positive (FP): Real events are classified as fake events, and False Negative (FN): Fake events are classified as real. Second there are user personal profile and behavioral features. These features are the two ways used to identify spammers Whether by using time-stamp of user's comment is frequent and unique than other normal users or if the user posts a redundant review and has no relation to domain of target. In this paper, We apply TF-IDF to extract the features of the contents in two languages models; mainly bi-gram and tri-gram. In both language models, we apply also the extended dataset after extracting the features representing the users behaviors.

5.3 Feature Engineering

Fake reviews are known to have other descriptive features related to behaviors of the reviewers during writing their reviews. In this paper, we consider some of these feature and their impact on the performance of the fake reviews detection process. We consider caps-count, punct-count, and emojis behavioral features. caps-count represents the total capital character a reviewer use when writing the review, punct-count represents the total number of punctuation that found in each review, and emojis counts the total number of emojis in each review. Also, we have used statistical analysis on reviewers' behaviours by applying "groupby" function, that gets the number of fake or real reviews by each reviewer that are written on a certain date and on each hotel. All these features are taken into consideration to see the effect of the users behaviors on the performance of the classifier

DOI: 10.48175/IJARSCT-13504
ISSN 2581-9429 IJARSCT

VI. CONCLUSION

This paper presented an extensive survey of the most notable works to date on machine learning-based fake review detection. The spam review detection using ML is designed for filtering the fake reviews. People write unworthy positive reviews about products to promote them. In some cases malicious negative reviews to other (competitive) products are given in order to damage their reputation. Some of these consists of non-reviews (e.g., ads and promotions) which contain no opinions about the product. We detecting the reviews that are not genuine or which are used to deviate the consumers opinion in a certain direction becomes even more difficult.

REFERENCES

[1]. Jindal, Nitin, and Bing Liu. "Opinion spam and analysis." Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008.

[2]. Lim, Ee-Peng, et al. "Detecting product review spammers using rating behaviors." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.

[3]. Algur, Siddu P., et al. "Conceptual level similarity measure based review spam detection." Signal and Image Processing (ICSIP), 2010 International Conference on. IEEE, 2010.

[4]. Feng, Song, et al. "Distributional Footprints of Deceptive Product Reviews." ICWSM12 (2012): 98-105.

[5]. Li, Wenbin, Ning Zhong, and Chunnian Liu. "Combining multiple email filters basedon multivariate
statistical analysis." Foundations of Intelligent Systems. Springer Berlin Heidelberg, 2006. 729-738. [6]. Liu, Bing, et al. "Partially supervised classi_cation of text documents." ICML. Vol. 2.2002.

[7]. Karimpour, Jaber, Ali A. Noroozi, and Somayeh Alizadeh. "Web Spam Detection by Learning from Small Labeled Samples." International Journal of Computer Applications 50.21, 2012.

[8]. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004.

[9]. Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT, 2011.

[10]. Duyu Tang, Bing Qin, and Ting Liu, Document modeling with gated recurrent neural network for sentiment classification. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 14221432, Lisbon, Portugal, 17-21 September, 2015.

[11]. Raghav, C., Nandan, L., Chaudhari, C., Shah, A. and Shingate, D.S., Sentiment Analysis for Business Intelligence Buildup-A Review Paper

DOI: 10.48175/IJARSCT-13504

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN SCIENCE, COMMUNICATION AND TECHNOLOGY
ISSN
2581-9429
IJARSCT