Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 2, No. 2, February : 2024

DECEPTIVE REVIEW DETECTION IN ONLINE SOCIAL NETWORKS

Nidhi A. Patel, Ph. D Research Scholar, Department of Computer Engineering, Gujarat Technological University, Ahmedabad, Gujarat, India; E-mail: nidhi.patel0051@gmail.com
Nirali Nanavati, Associate Professor, Department of Computer Engineering, Sarvajanik College of Engineering and Technology, Surat, Gujarat, India. E-mail: nirali.nanavati@scet.ac.in

Abstract: In today's e-commerce, online reviews are crucial for the decision-making process. Before choosing what, where, and when to buy, a sizable portion of the consumer base peruses product or store reviews. Deceptive reviews on internet sites have been increased significantly which can yield substantial profits to fraudsters. Positive reviews of the targeted product can attract more buyers and promote sales; negative reviews might drive away customers and reduce sales. These deceptive reviews are created purposefully to mislead prospective clients and tarnish their reputation. The goal of our work is to determine whether the review is factual or fraudulent. In this work, a Positive and Unlabelled (PU) machine learning based algorithm has been used. Our results show an outperformed existing PU based approach with respect to accuracy and other performance metrics.
Keywords: Online Social Media, Machine Learning, Deceptive Review, Spam Review, Semi-supervised Learning, Unsupervised Learning, Supervised Learning.

I. INTRODUCTION

Social media websites and online platforms have led to the widespread distribution of various forms of information (such as audio, video, and text) created entirely by users; this is known as user-generated content (UGC) [1]. In OSN, anyone can use social media content, even in the absence of a trustworthy external system of control. This indicates no methods for a priori verification or content production credibility [2]. Due to this issue, a lot of misinformation is being spread regarding the nature of spam and the harm it causes to users and businesses. This context's spam view detection aims to find fake comments, fake reviews, fake blogs, deceptions, misleading messages, and misleading public posts [3]. It can be challenging to distinguish deceptive reviews from other types of spam [4]. Thus, understanding the postings' context could be necessary to assess if a specific review is misleading [5].

Reviews are central to any comment, post, review, or tweet. Deceptive refers to any unsolicited or irrelevant information attached to these reviews for promotion, advertisement, information spread, or financial profit [6]. “Review spamming” refers to giving inaccurate or misleading information in reviews to mislead customers and affect product sales [18].

Compared to supervised and unsupervised, the proposed technique is typically based on semi-supervised modification techniques. However, there are other problems with supervised approaches that often only take labeled datasets, and unsupervised approaches take samples without labeled datasets. Therefore, the proposed solutions use a semi-supervised method. The proposed semi-supervised machine learning approach is to improve the classification with new dimensions (n-gram and word2vec features) feature vectors. Next, we evaluate the proposed approach using an existing method.

We review and analyze related work in the next area. The proposed method is explained in Section III. In Section IV, we present and discuss the experiment results, and the paper concludes with Section V.

RELATED WORK

A number of methods, including labeled (for example, supervised learning), unlabeled (for example, unsupervised learning), and partially labeled (for example, semi-supervised learning) data, have been proposed previously to detect deceptive reviews. A few papers on these methods are explained below.

Khurshid et al. [7] proposed an ensemble learning model using specific features to identify deceptive reviews with two tiers. Tier 1 used three classifiers (a library for SVM, Discriminative Multi-nominal NB, and J48), and Tier 2 used the LR classifier. The experimental findings show that the chi-squared feature with the ensemble model significantly enhances the performance of the suggested method. A supervised learning model based on unigram and bigram features models with two phases was presented by Mani et al. [8] to identify deceptive reviews. In the first, RF, SVM, and NB are used. Stacking and voting ensemble methods enhanced the classification model's performance in the second phase. An adaptation approach was presented by Sánchez-Junquera et al. [9] to identify deceptive reviews in cross-domain. The proposed framework frequently used co-occurring entropy to identify the domain features and then used a mismatch technique to conceal them. The proposed method struggled to identify deceptive reviews in cross-domain, according to the results of the standard dataset using NB classifiers. In order to examine review inconsistency based on various features (language, content, and rating) in detecting deceptive reviews, Shan et al. [10] established a framework. To determine if a review is real or deceptive, the retrieved features are input into different ML classifiers (NB, SVM, MLP, and RF). The experimental results indicate that features related to review inconsistency can enhance the efficacy of detecting deceptive reviews. In order to identify deceptive reviews, Goyal et al. [19] used a hybrid approach that involved four stages: First, preprocessing the data using the Natural Language Toolkit (NLTK), extracting informative features using Term frequency-Inverse Document Frequency (TF-IDF) parameters, sentiment analysis scores, and syntactic patterns. The Bernoulli Naïve Bayes (BNB), Gaussian Naïve Bayes (GNB), and Multinomial Naïve Bayes (MNB) algorithms were used to train the model on these generated features. Four preprocessing steps were utilised by Asaad et al. [20], tokenization, normalisation, stop word removal, and stemming. TF-IDF approaches were then used to extract features. Three machine learning techniques used by the authors for the classification: stochastic gradient descent, support vector classifier, and Xgboost.

An unsupervised topic sentiment model was presented by Dong et al. [11] to detect deceptive reviews. The four layers of the proposed model were word, subject, document, and sentiment. The authors improved the LDA framework used to extract subject sentiment from reviews by extracting topic information from documents. SVM and RF classifiers are given sentiment and topic features. The Gibbs sampling approach was used to derive the probability distribution between words and topics, as well as between topics and sentiment. A technique for identifying a set of deceptive reviews based on nominated topics has been proposed by Li et al. [12]. The three stages of the proposed model are as follows: first, they define the equivalent groups and their target topics, then use the K-means algorithm to cluster reviews. Lastly, they labeled the suspicious group deceptive using time burstiness and content duplication. Li. et al. [21] propose two models that fit reviews across JD.com and TMALL.com using aspect-oriented semantic mining. Reviews are grouped into spam suspect and benign groups based on the quantifiable correlation levels to product metadata and nominated topic.

Yilmaz et al. [13] proposed a semi-supervised learning framework (SPR2EP) for detecting deceptive reviews using reviewer item network attributes and text content to identify fraudulent reviews. Two learning algorithms were used, namely node2vec and Doc2vec. These representations are then loaded into a logistic regression model to determine whether or not reviews are spam. Tian et al. [14], a semi-supervised algorithm known as ‘‘Ramp One-Class SVM’’ was applied to detect deceptive reviews.

A. Research Gap

As per the literature, the problems are inaccurate predictions [7], slow convergence [8], computationally expensive [9], time-consuming and resource-intensive [10], and computational complexity [11] [13] for correctly identifying deceptive review detection. To overcome this, we have proposed a detection model with a good learning paradigm.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 2, No. 2, February : 2024

II. PROPOSED WORK

A. Data set Description

The data set that we have utilized includes 1600 reviews, 800 of which are deceptive reviews and 800 genuine reviews. Of these, 400 reviews have negative sentiment polarities, and 400 show positive sentiment polarities. Positive opinion reviews are a combination of deceptive and truthful reviews. We have collected the dataset from Ott et al. [15] [17] for the review spam detection.

B. Data Preprocessing

The main standard preprocessing steps are considered in this paper including: tokenization and punctuation marks removal. The tokenization is separating the text into a small number of words or sentences. The crucial preprocessing step is punctuation mark removal, which divides the text into paragraphs, sentences, and phrases. Word2Vec is a method for creating word embeddings. The purpose of word2vec is to group the vectors of related words together in vectorspace.

C. Feature Engineering

Feature engineering is the process of creating or extracting features from data. Our proposed approach used a "bag of words" (BOW) strategy. In this approach, individual word groups are found in the text. These fractures, known as n-grams, are created by choosing a continuous word from a specific sequence. In the proposed approach, we have used bigram and trigram (n = 2 and 3) and word2vec features and compared the results with the existing approach [17]. The results are shown in section IV.

D. Proposed Algorithm

The below sequential steps show the pseudocode of the proposed PU Learning approach.

PU-Learning for Spam Review Detection

1 Preprocessing: Tokenization and Punctuation Marks Removal
2 Feature Engineering: N-gram and Word2Vec
3 $i \leftarrow 1$;
4 $|W_0| \leftarrow |U_1|$;
5 $|W_1| \leftarrow |U_1|$;
6 while $|W_i| \leq |W_i - 1|$ do
7 $Ci \leftarrow$ Generate Classifier(P, U_i);
8 $U_i^L \leftarrow Ci\ (U_i)$;
9 $W_i \leftarrow$ Extract Positives(U_i^L);
10 $U_{i+1} \leftarrow U_i - W_i$;
11 $i \leftarrow i + 1$;
12 Return Classifier Ci

The proposed approach is based on the PU learning method [16]. It is an iterative procedure where unlabeled datasets are treated as negative classes in this approach. Next, we trained various classifiers using positive cases. Here, six classifiers have been used. These include DT, NB, SVM, KNN, RF, and LR classifiers. After, these classifiers to classify unlabeled datasets. All positive examples are removed from instances of unlabeled data, and the remaining instances are treated as negative instances for the next iteration. This process is repeated until the stop condition is fulfilled. The flowchart of the proposed methodology is shown in Figure 1.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 2, No. 2, February : 2024

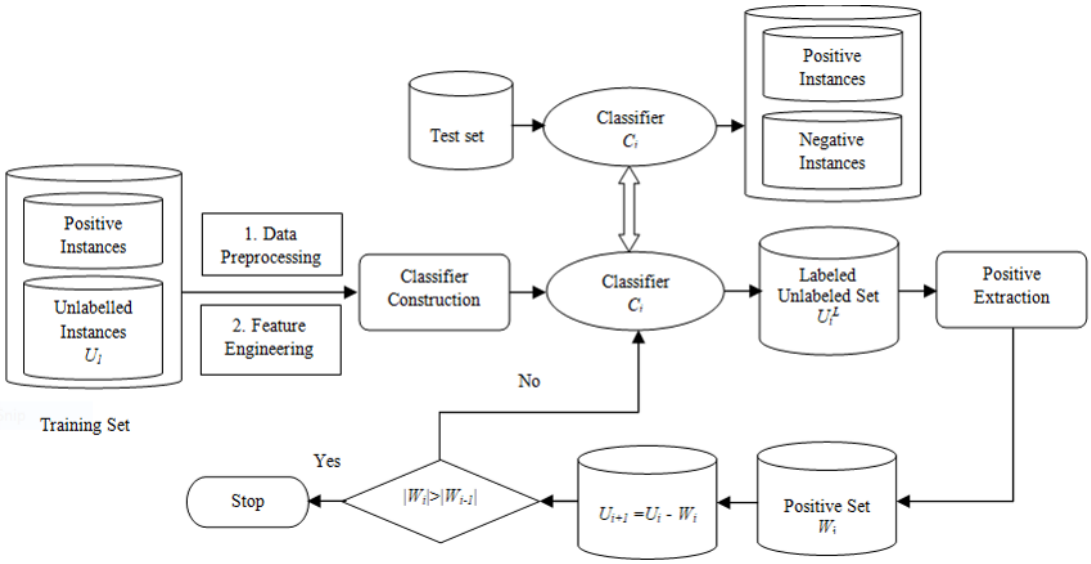

Figure 1. Flowchart of the Proposed Methodology

Table 1. The result of different classifiers using 40 deceptive reviews as training and 520 unlabeled reviews with feature method of N-gram with Bi-gram and Tri-gram and Word2Vec.

	Proposed Approach with N-gram								**Proposed Approach with Word2Vec**			
	Bi-gram				**Tri-gram**				**Word2Vec**			
Classifier	**A (%)**	**P**	**R**	**F**	**A (%)**	**P**	**R**	**F**	**A (%)**	**P**	**R**	**F**
DT	58.36	0.59	0.52	0.55	61.35	0.68	0.63	0.65	63.72	0.69	0.65	0.67
NB	38.53	0.45	0.36	0.40	40.13	0.45	0.36	0.40	44.36	0.48	0.39	0.43
SVM	55.23	0.45	0.54	0.49	55.26	0.54	0.51	0.52	56.92	0.54	0.52	0.53
KNN	65.32	0.67	0.61	0.64	65.14	0.69	0.71	0.70	69.25	0.72	0.70	0.71
RF	56.27	0.49	0.52	0.50	57.32	0.62	0.56	0.59	60.01	0.64	0.61	0.62
LR	62.36	0.59	0.61	0.60	62.84	0.72	0.74	0.73	64.25	0.74	0.73	0.73

Table 2. The result of different classifiers using 80 deceptive reviews as training and 520 unlabeled reviews with feature method of N-gram with Bi-gram and Tri-gram and Word2Vec.

	Proposed Approach with N-gram								**Proposed Approach with Word2Vec**			
	Bi-gram				**Tri-gram**				**Word2Vec**			
Classifier	**A (%)**	**P**	**R**	**F**	**A (%)**	**P**	**R**	**F**	**A (%)**	**P**	**R**	**F**
DT	71.36	0.73	0.68	0.70	72.34	0.74	0.71	0.72	74.63	0.78	0.71	0.74
NB	55.51	0.58	0.52	0.55	56.24	0.54	0.51	0.52	56.93	0.61	0.52	0.56
SVM	73.46	0.74	0.66	0.70	72.35	0.76	0.78	0.77	75.36	0.79	0.80	0.79
KNN	76.25	0.78	0.79	0.78	76.89	0.84	0.79	0.81	79.25	0.81	0.78	0.79
RF	65.24	0.68	0.67	0.67	67.27	0.65	0.64	0.64	69.25	0.67	0.62	0.64
LR	77.25	0.79	0.76	0.77	77.26	0.73	0.75	0.74	81.53	0.83	0.76	0.79

Table 3. The result of different classifiers using 120 deceptive reviews as training and 520 unlabeled reviews with feature method of N-gram with Bi-gram and Tri-gram and Word2Vec

	Proposed Approach with N-gram								**Proposed Approach with Word2Vec**			
	Bi-gram				**Tri-gram**				**Word2Vec**			
Classifier	**A (%)**	**P**	**R**	**F**	**A (%)**	**P**	**R**	**F**	**A (%)**	**P**	**R**	**F**
DT	48.25	0.50	0.46	0.48	48.82	0.53	0.49	0.51	51.36	0.54	0.51	0.52
NB	55.27	0.58	0.57	0.57	55.73	0.53	0.52	0.52	55.91	0.58	0.54	0.56
SVM	61.43	0.68	0.59	0.63	63.43	0.63	0.61	0.62	64.25	0.64	0.61	0.62
KNN	62.37	0.56	0.55	0.55	73.25	0.71	0.75	0.73	76.92	0.73	0.71	0.72
RF	48.62	0.45	0.43	0.44	55.91	0.52	0.54	0.53	58.04	0.59	0.56	0.57
LR	75.26	0.72	0.71	0.71	79.91	0.91	0.76	0.83	**82.81**	0.92	0.81	0.86

Table 4. The result of different classifiers using 120 deceptive reviews as training and 520 unlabeled reviews with existing and proposed with Word2Vec feature method

	Existing [17]				**Proposed (Word2Vec)**			
Classifier	**A (%)**	**P**	**R**	**F**	**A (%)**	**P**	**R**	**F**
DT	45.31	50.00	45.71	47.76	51.36	0.54	0.51	0.52
NB	54.68	34.37	57.89	43.13	55.91	0.58	0.54	0.56
SVM	60.93	90.92	56.86	69.87	64.25	0.64	0.61	0.62
KNN	60.93	71.87	58.97	64.78	76.92	0.73	0.71	0.72
RF	46.87	56.25	47.36	51.42	58.04	0.59	0.56	0.57
LR	73.43	68.75	75.86	72.13	**82.81**	0.92	0.81	0.86

III. Experimental Results and Discussion

Our results with the semi-supervised learning method, we used in our experiments yielded the following results: As mentioned in section 3 for the data set, we have implemented our model in Python. Tables 1, 2, and 3 display the results for different training sets. For building test data, we randomly selected 160 opinion reviews with a combination of deceptive and truthful reviews. The 640 opinion reviews have been applied to training sets of various sizes. We consist of 40, 80, and 120 deceptive opinion instances, respectively. We have used 520 unlabeled instances in all the cases as per existing [17]. We utilized the following six classifiers: 1) Decision Tree (DT), 2) Naive Bayes (NB), 3) Support Vector Machine (SVM), 4) K-Nearest Neighbor (KNN), 5) Random Forest (RF), and 6) Logistic Regression (LR). We considered the accuracy (A), precision (P), recall (R), and f-score (F) parameters for evaluation and compared the results.

Tables 1, 2, and 3 compare the proposed with bi-gram, tri-gram of n-gram, and word2vec features. Out of all the results, word2vec got better results. Table 4 shows the result of 120 deceptive reviews as training and 520 unlabeled reviews with the existing and proposed with word2vec feature method.

A. Discussion

The highest level of accuracy we have achieved is 82.81 % when using 120 deceptive opinion reviews as training and 520 unlabeled opinion reviews using logistic regression. The logistic regression works on containing maximum likelihood estimation and using a softmax classifier that divides multiple classes of data and works well with the textual dataset.

IV. Conclusion and Future scope

In this work, a PU based machine learning algorithm was applied using preprocessing and feature engineering for better prediction accuracy. For preprocessing, we simply applied tokenization and

removed punctuation marks including white spaces. For feature engineering, we evaluated our approach using n-gram (namely bigram and trigram) and word2vec methods and observed that word2vec gives comparatively good results. We experimented our approach with different supervised machine learning algorithms namely decision tree, naive bayes, support vector machine, k-nearest neighbor, random forest, and logistic regression. From the results, we found that logistic regression based approach outperforms existing PU based approach. In the future, the same work can be extended with more features with other machine learning algorithms.

REFERENCES

[1] Petrescu Maria, Ajjan Haya, & Harrison, Dana. L. Harrison. 2023. "Man vs machine – detecting deception in online reviews". *Journal of Business Research*, *154*, 113346. https://doi.org/10.1016/j.jbusres.2022.113346.

[2] Zhang Yubao, Wang Haining, & Stavrou Angelos. 2023. "A multiview clustering framework for detecting deceptive reviews" *Journal of Computer Security*, 1–22. https://doi.org/10.3233/jcs-220001.

[3] Jacob Minu Jacob & Selvi Rajendran. 2022. "Deceptive Product Review Identification Framework using opinion mining and machine learning" *Mobile Radio Communications and 5G Networks*, 57–72. https://doi.org/10.1007/978-981-16-7018-3_4.

[4] Lee Minwoo, Song Young, Li Lin, Lee Kyung, and Sung-Byung Yang. 2022. "Detecting fake reviews with supervised machine learning algorithms" *The Service Industries Journal*, *42*(13–14), 1101–1121. https://doi.org/10.1080/02642069.2022.2054996.

[5] Nagi Alsubari Saleh, Deshmukh Sachin, Abdullah Alqarni Ahmed, Alsharif Nizar, Aldhyani, Theyazn, Waselallah Alsaade Fawaz, and Khalaf Osamah. 2022. "Data Analytics for the identification of fake reviews using supervised learning" *Computers, Materials & Continua*, *70*(2), 3189–3204. https://doi.org/10.32604/cmc.2022.019625.

[6] Alawadh Husam, Alabrah Amerah, Meraj Talha, and Rauf Hafiz. 2023. "Semantic features-based discourse analysis using deceptive and real text reviews", *Information*, *14*(1), 34. https://doi.org/10.3390/info14010034.

[7] Faisal Khurshid, Yan Zhu, Zhuang Xu, Mushtaq Ahmad, and Muqeet Ahmad. 2019. "Enactment of Ensemble Learning for Review Spam Detection on selected features", *International Journal of Computational Intelligence Systems*, *12*(1), 387. https://doi.org/10.2991/ijcis.2019.125905655.

[8] Shwet Mani, Sneha Kumari, Ayushi Jain and Prabhat Kumar. 2018, ''Spam review detection using ensemble machine learning,'' *Machine Learning and Data Mining in Pattern Recognition*, 198–209. https://doi.org/10.1007/978-3-319-96133-0_15.

[9] Javier Sánchez-Junquera, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez, Paolo Rosso, and Efstathios Stamatatos. 2020. ''Masking domain-specific information for cross-domain deception detection,'' *Pattern Recognition Letters*, *135*, 122–130. https://doi.org/10.1016/j.patrec.2020.04.020.

[10] Guohou Shan, Lina Zhou, and Dongsong Zhang. 2021. ''From conflicts and confusion to doubts: Examining review inconsistency for fake review detection,'' *Decision Support Systems*, *144*, 113513. https://doi.org/10.1016/j.dss.2021.113513.

[11] Lu-yu Dong, Shu-juan Ji, Chun-jin Zhang, Qi Zhang, DicksonK.W. Chiu, Li-qing Qiu, and Da Li. 2018, ''An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews,'' *Expert Systems with Applications*, *114*, 210–223. https://doi.org/10.1016/j.eswa.2018.07.005.

[12] Jiandun Li, Pin Lv, Wei Xiao, Liu Yang, and Pengpeng Zhang. 2021. ''Exploring groups of opinion spam using sentiment analysis guided by nominated topics,'' *Expert Systems with Applications*, *171*, 114585. https://doi.org/10.1016/j.eswa.2021.114585.

[13] Cennet M. Yilmaz and Ahmet O. Durahim. 2018. ''SPR2EP: A semi-supervised spam review detection framework,'' *2018 IEEE/ACM International Conference on Advances in Social Networks*

Analysis and Mining (ASONAM), 306-313, https://doi.org/10.1109/asonam.2018.8508314.
[14] Yingjie Tian, Mahboubeh Mirzabagheri, Peyman Tirandazi, and Seyed Bamakan. 2020. ‘‘A nonconvex semi-supervised approach to opinion spam detection by rampone class SVM,’’ *Information Processing & Management*, *57*(6), 102381. https://doi.org/10.1016/j.ipm.2020.102381.
[15] Ott Myle,Yejin Choi, Claire Cardie and Jeffrey Hancock. 2011, “Finding deceptive opinion spam by any stretch of the imagination”, *In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics,* Volume 1, https://doi.org/10.48550/arXiv.1107.4557.
[16] Hernndez Donato, Rafael Guzman-Cabrera and Manuel Montes. 2013. “Using PU-learning to detect deceptive opinion spam” *Proc. of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 38-45.
[17] Narayan Rohit, Rout Jitendra Kumar and Jena Sanjay Kumar. 2017. “Review spam detection using semi-supervised technique”, *Advances in Intelligent Systems and Computing*, 281–286. https://doi.org/10.1007/978-981-10-3376-6_31.
[18] Ott Myle, Claire Cardie, and Jeffrey T. Hancock. 2013. “Negative Deceptive Opinion Spam” *Proceedings of NAACL-HLT, Association for Computational Linguistics*, 497–501.
[19] Goyal, N. K., Pal, A., Keswani, B., Goyal, D., & Gupta, M. K. (2023). A novel hybrid feature extraction technique and Spam Review Detection Using Ensemble Machine Learning algorithm by web scrapping. *Indian Journal Of Science And Technology*, *16*(29), 2261–2268. https://doi.org/10.17485/ijst/v16i29.1500.
[20] Asaad, W. H., Allami, R., & Ali, Y. H. (2023). Fake review detection using machine learning. *Revue d’Intelligence Artificielle*, *37*(5). https://doi.org/10.18280/ria.370507.
[21] Li, J., Li, N., Yang, L., & Zhang, P. (2022). Identifying review spam with an unsupervised approach based on topic abuse. *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*. https://doi.org/10.1145/3532213.3532265.