I.J. Information Engineering and Electronic Business, 2020, 1, 1-7
Published Online February 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2020.01.01

Deceptive Opinion Detection Using Machine Learning Techniques

Naznin Sultana[1, 2]
[1]Department of Information Technology, Malaysia University of Science & Technology, Petaling Jaya, Malaysia
[2]Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh
Email: nazninsultana60@gmail.com

Prof. Sellappan Palaniappan
Department of Information Technology, Malaysia University of Science & Technology, Petaling Jaya, Malaysia
Email: sell@must.edu.my

Received: 15 June 2019; Accepted: 28 October 2019; Published: 08 February 2020

Abstract—Nowadays, online reviews have become a valuable resource for customer decision making before purchasing a product. Research shows that most of the people look at online reviews before purchasing any product. So, customers reviews are now become a crucial part of doing business online. Since review can either promote or demote a product or a service, so buying and selling fake reviews turns into a profitable business for some people now a days. In the past few years, deceptive review detection has attracted significant attention from both the industrial organizations and academic communities. However, the issue remains to be a challenging problem due to the lack of labeled dataset for supervised learning and evaluation. Also, study shows that both the state of the art computational approaches and human readers acquire an error rate of about 35% to 48% in identifying fake reviews. This study thoroughly investigated and analyzed customers' online reviews for deception detection using different supervised machine learning methods and proposes a machine learning model using stochastic gradient descent algorithm for the detection of spam review. To reduce bias and variance, bagging and boosting approach was integrated into the model. Furthermore, to select the most appropriate features in the feature selection step, some rules using regular expression were also generated. Experiments on hotel review dataset demonstrate the effectiveness of the proposed approach.

Index Terms—Natural Language Processing, Spam Review, Opinion Mining, Ensemble Learning, Machine Learning.

I. Introduction

With the widespread use of internet and web technology e-commerce web sites and online marketplace plays an important role to reach wider customers in a very short time. So the number of online reviews by customer is increasing as well. These e-commerce web sites consist of an enormous amount of data about customers' and consumers' product experiences and their opinion about the product. This information often acts as an indicator of the quality of products and thus has a great impact on purchasing decisions of consumers, retailers, and manufacturers. In these online platform customers usually express their opinion as text reviews which become ubiquitous and assist buyers in making purchase decisions. So customers reviews are now become a crucial part of doing business online.

In order to boost up and raise their businesses, business owners, retailers and service providers often expects and ask their potential customers to provide some positive comments about their products/services they have bought/used. While online reviews can be helpful in most of the cases, however, sometimes these reviews can be hazardous for both the retailer and purchaser when the reviews are fake. Business owners sometimes hire third party via the internet or from some other source to write fake reviews, either paid or unpaid basis. They may write good reviews about their merchandise or bad reviews towards their competitors' products/services. These fake reviews are called deceptive opinion or review spam which has a great impact in e-marketplace nowadays. Deceptive opinion has a negative impact on business due to the loss in customers and consumers trust. As review spam becomes a prevalent and widespread problem, so the development of some methods to help businesses and customer to identify truthful reviews from fake ones are the most needed task.

Review spam is somewhat related to the web or email spam but since it deals with false opinions, so it is much harder to detect than the other two spam. So the existing methods for detecting web spam and email spam [1,2,3], is not suitable for review spam. Spam reviews can be of different types. According to literature [4], opinion spam can be categorized into two types:

- Deceptive opinion spam
- Disruptive opinion spam

Deceptive opinion spam can be further divided into

positive opinion spam and negative opinion spam. Generally, reviews containing only advertisements or random texts are not usually considered as disruptive opinion spam and are much easier to detect by manual inspection. So, current literature on opinion spam detection mainly focused on detection of deceptive spam, which is more harmful and much harder to detect because these reviews are fabricated in such a way to look like authentic ones. This paper focuses on deceptive opinion spam and its detection in detail.

Though there are a good number of machine learning algorithms for the detection of true and fake reviews but most of the research mainly focused on using supervised learning methods based on three basic algorithms: Naïve Bayes, Logistic Regression and Support Vector Machine. Also there has been found few works on review spam detection as compared to email and web spam detection. Therefore, to investigate the effectiveness of different machine learning algorithms in combination with ensemble learning and rule based feature selection approach for deceptive opinion detection is the main objective of this study.

For detecting review spam we have considered review content from the dataset which represents the actual text written by the reviewer. Some linguistic features are extracted from these review content representing fraudulent behavior. We first analyzed the model performance using six popular and widely used supervised machine learning algorithms. Ensemble approach using bagging and boosting were also applied on two different base classifiers. We analyzed the performance of these algorithms in different perspectives and finally, we came up with a conclusion about the prediction capability of the selected algorithms with the help of some evaluation matrices. The results of our investigation can be used in a variety of large scale textual data processing systems for selecting the model structure and to choose the optimal algorithm based on the nature of the dataset. In addition, our findings will also help data analysts to predict the data to support knowledge gathering and decision support system. The rest of the paper is organized in the following manner: Section 2 provides related works from literature; Section 3 describes the datasets and experimental setup of our model; analysis of result is discussed in Section 4 and Section 5 concludes the paper with the future extension of this work.

II. Related Works

In literature most of the work in spam detection has been done in detecting web spam and email spam. Recently, some of the researchers have started working on opinion spam as well. In [5] authors analyzed the effect of supervised, unsupervised and semi-supervised learning along with feature engineering approach. For review spam detection this paper discussed the review centric features such as bag of words in combination with tf-idf, parts of speech, syntactic and stylometric features. In the absence of gold-standard dataset, Jindal and Liu [6] used product review data and train their model using extraction of features from review text, reviewer, and product to detect untruthful (duplicate opinions) and truthful (non-duplicate opinions) reviews. Using logistic regression they achieved an AUC score of 78%. Researchers in [7] proposed a strategy based on user centric and user behavior driven approach for detecting deceptive opinion on Amazon review dataset. They suggested a model based on patterns of review content and ratings to define four different spamming behaviors, i.e. targeting product; targeting group; general rating deviation; and early rating deviation. They showed that their proposed model outperform other baseline method based on helpfulness votes alone. G Feietal in [8] exploits the burstiness nature of reviews to identify review spammers since bursts of reviews can be either due to sudden popularity of products or spam attacks. They model reviewers and their co-occurrence in bursts as a Markov Random Field (MRF), and employ the Loopy Belief Propagation (LBP) method to infer whether a reviewer is a spammer or not in the graph and achieved an accuracy of 77.6%. As like as in [8], researchers in [9] considered that spam attacks are usually bursty and either positively or negatively correlated to the rating. They proposed a model to detect such attacks via unusually correlated temporal patterns. Their proposed hierarchical algorithm robustly detect the time windows where such attacks are likely to be happened. Experimental results showed that the proposed method is effective in detecting singleton review attacks. Measurement of text similarity based on Kullback-Leibler was proposed by authors in [10]. They used SVM for model prediction purpose and obtained similar results as in [6]. Authors in [11] used a content-based approach and achieved almost 90% accuracy using SVM and Naïve Bays algorithm. For feature extraction, they used POS, n-gram and LIWC output and proved that SVM outperformed Naïve Bayes by using these features. In [12] the same authors addressed review spam only for negative reviews. They collected 400 truthful negative reviews from six different websites and 400 negative fake reviews from AMT (Amazon Mechanical Turk). They used n-grams features and SVM algorithm. The accuracy they obtained was 86%. Researchers in [13] proposed a model for synthesizing deceptive reviews from true ones and claimed that even the best algorithms in deceptive detection have an error rate higher than 30%. Researchers in [14] used behavioral features on a dataset from Yelp. Some meta-data as the maximum number of reviews, review length, reviewer deviation, positive review percentage, and maximum content similarity were selected as features in their model. Using bi-gram and SVM with 5 fold cross validation they achieved 64.4% accuracy. Only behavior features (BF) yielded 83.2%, while the combination of bigrams and BF the accuracy they found was 84.8%. These results showed that methods using behavior features achieved much better results than content-based methods on Yelp dataset. They also tested the effect of excluding one or more features from their model and the impact in terms of accuracy. Singleton review and group review (spammer group) are discussed in [1].

III. Dataset Description And Experiment Setup

A. Dataset Description

The dataset used in our experiment was collected from online source [11,12], which consists of total 1600 reviews from TripAdvisor, Yelp.com, Expedia, Hotels.com, Priceline and Amazon Mechanical Turk. There are two types of reviews in the dataset, i.e. reviews representing positive sentiment and reviews representing negative sentiment. The description of the dataset is shown in Table 1:

Table 1. Dataset Description

Review Type	No. of Positive Reviews	No. of Negative Reviews
Truthful	400	400
Deceptive	400	400

All the data are hotel reviews from 20 hotels with each review has the following attributes:

- A unique ID
- Hotel name about which review has been written
- Review content
- Review polarity i.e. whether the review portrays positive or negative sentiment.

The data corpus consists of 80 reviews for each of the 20 hotels. Among these 80 reviews 40 reviews were truthful and 40 deceptive. In each of those categories 20 reviews were positive and 20 negative. The dataset was made well balanced by keeping each hotel an equal numbers of reviews in all categories.

B. Preparation of Review Text

This stage is concerned with the preparation of review text to extract features and fed into the machine learning model. Following operations were performed as the data preparation tasks:

- Tokenizing each word of the text and giving an integer id for each possible token by using punctuation or white space as token separators.
- Removing all stop words such as 'a' and 'the' (Stop word corpus was taken from the NLTK website. Stop words 'a' and 'the' are frequently used in any text, but they do not actually carry any specific information required to train the model.

- Converting all the capital letters to a lower case.
- Pulling out numeric values from review text.
- Lemmatizing to group together the different forms of the same word.

C. Experimental Setup

Our experiment mainly consists of two steps: duplicate detection and spam classification. However, we performed some other tweaking operations on the review text to optimize the accuracy of the model. Fig.1 shows the workflow model for spam review detection.

1) Duplicates detection

In review spam detection lack of standard datasets makes it difficult to compare results from different studies. As our dataset has few features we have considered review text as the only features for deception detection in training and testing phase. So, our first task in review spam detection is to find out duplicate or near duplicate reviews from the review text since it has been found that many of the duplicate reviews are clearly spam. For example, same user id or different user id write one review and duplicate that reviews on the same product or different products. In our experiment, we have used cosine similarity measure from python library to detect duplicate review for similarity score greater than 0.9.

However, spammers often copy genuine reviews or write a review that looks like a genuine review. So using this technique may result in both genuine and fake reviews to be identified as deceptive. So, besides similarity measures, below are some other techniques we followed in detecting spam from review text:

a) Bag of Words

This approach considers words or sequence of words in review texts as features where sequences of words are called n-grams (n = 1, 2, 3… is the no. of words in a sequence). In our experiment we have used 2-gram.

b) Term Frequency

It is a scoring scheme used in information retrieval. TF-IDF measures how relevant a term is in a given document. The idea is, if a word occurs multiple times in a document, then it should be more meaningful than other words that appear fewer times.

c) POS tags

A POS tag (or part-of-speech tag) is the process of marking up a word in a text corresponding to a particular part of speech based on its definition and context. This process assigns special tag to each word in a text either as noun, verbs, adverbs etc. based on its definition and relationships with adjacent words.

d) Stylometric features

This technique tries to capture reviewers writing style. Stylometric features include number of punctuation marks used in reviews, number of emotional words, average number of words per sentence, no. of characters in sentences etc.

e) Semantic

This feature deals with the meaning and interpretation of words, signs and sentence structure. They include synonyms and similar phrases. The idea of using this feature is that spammers sometimes replaces some words with their synonyms keeping the review same while making it harder to detect duplicate reviews.

For feature selection we generated the following chinking and chunking pattern using regular expression (Regex) in python as in (1) and applied on the POS tagged words to filter out the most appropriate features based on the corresponding pattern during feature extraction phase:

$$\begin{cases} R1:\{<JJ>+<NN\mid NNS>*\} \\ R2:\{<RB\mid RBR\mid RBS>*<JJ>*\}<NN\mid NNS>+\{ \\ R3:\{<JJ>*<JJ>*\}<NN\mid NNS>+\{ \\ R4:\{<NN\mid NNS>*<JJ>*\}<NN\mid NNS>+\{ \\ R5:\{<RB\mid RBR\mid RBS>*<VB\mid VBN\mid VBD\mid VBG>*\} \end{cases} \quad (1)$$

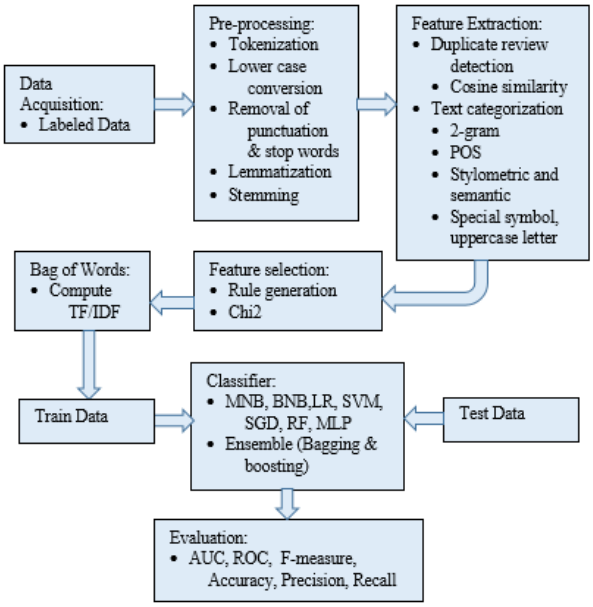


Fig.1. Workflow model for review processing

2) Spam classification

For classification of spam we have used two-class classification i.e. spam (deceptive) and non-spam (truthful). The ratio of train test split for the dataset was set as 70:30. The experiment was conducted using six different supervised machine learning algorithms, i.e. logistic regression, Naïve Bayes (Multinomial and Bernouli), Support Vector Machine(SVM), Stochastic Gradient Descent (SGD), Random Forest and Multilevel Perceptron (MLP). Ensemble learning (bagging and boosting) was employed on SVM and SGD base classifiers for the classification task.

IV. Analysis of Results

We used python 3.7.3 with windows operating system to implement our machine learning model. Some of the evaluation matrices like AUC-ROC, precision, recall, confusion matrix, correlation coefficient, F- measure and accuracy are calculated to observe the model performance. Table 2 summarizes the result of our experiment for the different classifiers used in our model. We performed 5-fold cross-validation on our dataset and we got an average AUC score of 88%, which is quite high considering that many non-spam text reviews are actually spam and thus have similar probabilities as spam reviews.

We empirically observed that Stochastic Gradient Descent (SGD) with its bagged and boosted approach has produced the best accuracy which is 78.3%. SVM also produces a similar but a bit less accuracy, F1 score and AUC than SGD. Fig. 2 illustrates the receiver operating characteristics (ROC) curve of different classifiers for a cross- validation score of 5. A ROC is a graphical presentation showing the performance of a classifier at different classification thresholds. This curve plots the ratio of true positive rate (TPR) vs. false positive rate (FPR).

Table 2. Performance Analysis of Different Classifiers on Hotel Review Dataset

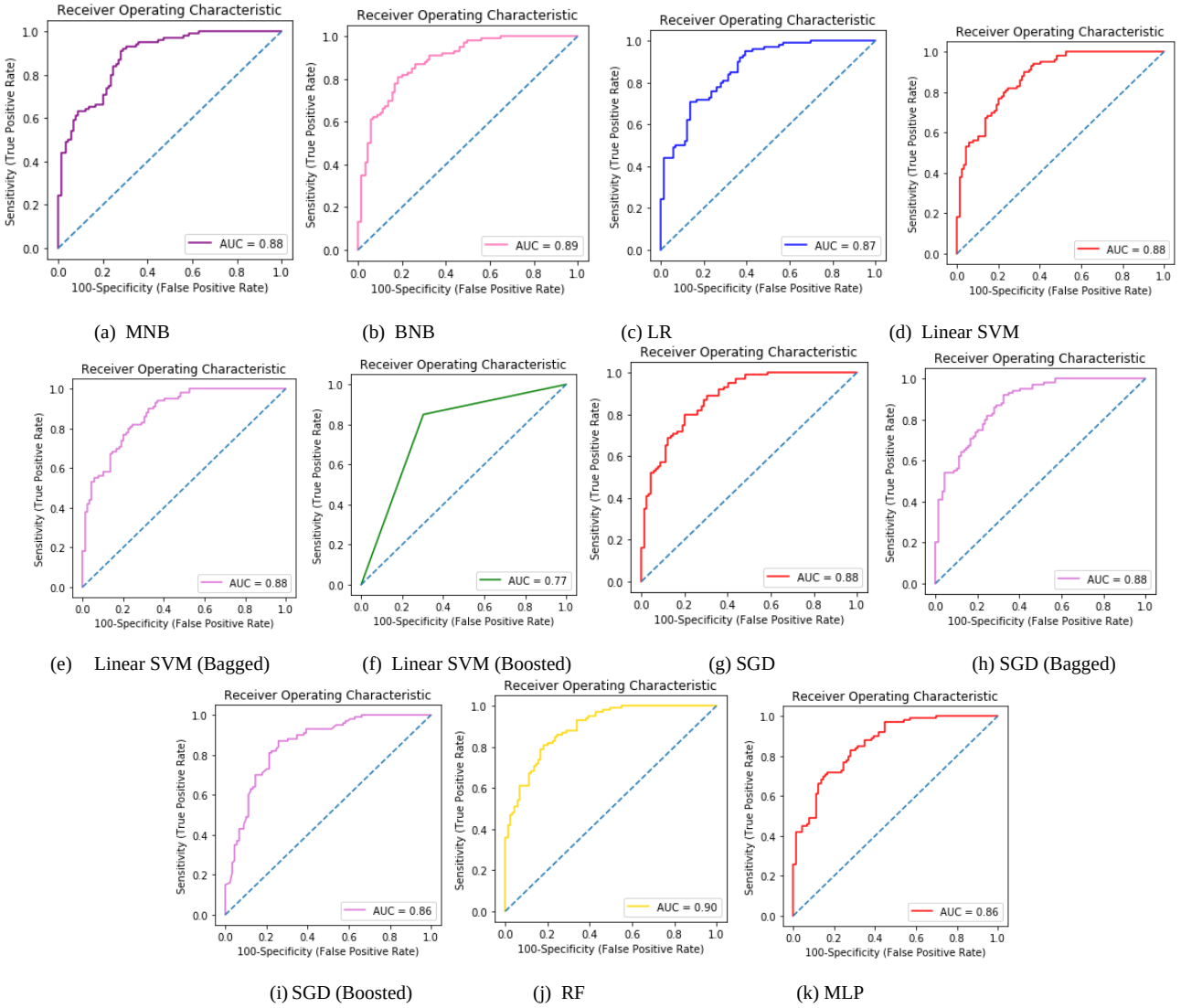
Classifier	Log Loss	Matthews Correlation Coefficient	F1 Score	AUC Score	Accuracy
MNB	0.60207	0.47855	0.77821	0.88269	69.84%
BNB	4.45316	0.20788	0.70921	0.88550	56.61%
LR	2.40480	0.56393	0.81081	0.86573	77.78%
Linear SVM	3.27716	0.55530	0.80188	0.87797	77.77%
Linear SVM (Bagged)	3.27716	0.55530	0.80188	0.87797	77.77%
Linear SVM (Boosted)	7.67539	0.55530	0.80188	0.77331	77.77%
SGD	3.68649	0.55351	0.79611	0.88101	77.77%
SGD (Bagged)	3.25105	0.56484	0.80382	0.87865	78.30%
SGD (Boosted)	2.41226	0.55411	0.80717	0.85544	78.30%
RF	0.61829	0.57649	0.81481	0.89617	76.19%
MLP	1.69727	0.51432	0.78703	0.86337	75.66%

True Positive Rate (TPR) is also called recall and is defined as:

TPR=True Positive / (True Positive +False Negative)

False Positive Rate (FPR) is called precision and is defined as:

FPR=False Positive / (False Positive +True Negative)



(a) MNB (b) BNB (c) LR (d) Linear SVM

(e) Linear SVM (Bagged) (f) Linear SVM (Boosted) (g) SGD (h) SGD (Bagged)

(i) SGD (Boosted) (j) RF (k) MLP

Fig.2. (a)-(k) ROC curve of different classifiers

Fig. 3 and Fig. 4 illustrates the F1 score and accuracy of different classifiers used in this experiment. From the figures it has been found that the accuracy of Logistic Regression, Linear SVM and Stochastic Gradient Descent is almost similar. However, SGD bagged and boosted has achieved 1.53% higher average classification accuracy in comparison with the other two for this dataset. All other classifier has produced less accuracy than SGD, so inference can be made that SGD is more stable and less distributed among all other machine learning models for review spam classification from text review. So, according to this study we can conclude that SGD with its bagged and boosted form has a big potential to improve the performance of spam review detection model. However, as the dataset we have used in our experiment was only the hotel reviews, so it does not make sense to combine or compare it with other reviews for other categories of products.

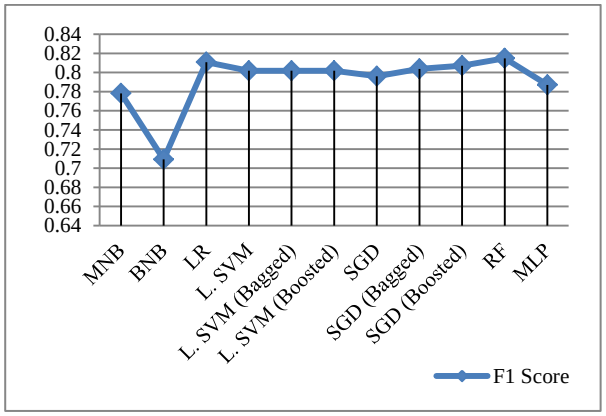


Fig.3. F1 Score of different classifiers

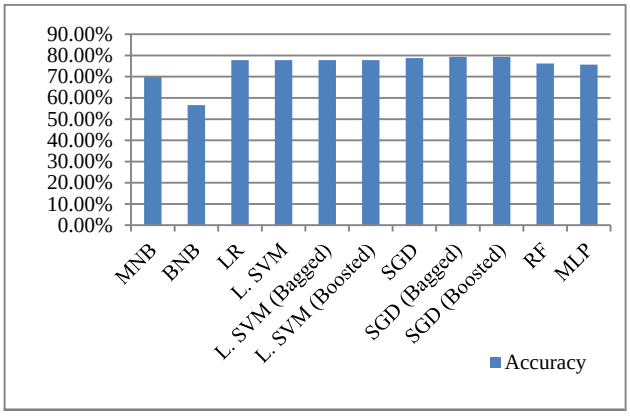


Fig.4. Accuracy of different classifiers

V. Conclusion

Now a day's deception detection draws the attention of many researchers due to the impact of deceptive opinion both on consumer behavior and customer purchase decision. So, it has a great impact in business and academia as well. This study compares the different machine learning models that have been proposed so far for deception detection through opinion mining. Though supervised learning is the most frequently used approach in this case, however, obtaining labeled dataset with lots of features is almost rare. Also, manual labeling of reviews has also found to be poor in accuracy. Due to this fact many researchers do their studies using synthetically generated small sized labeled datasets. But recent research suggests that experiments should use real-world data since models that use synthetically generated dataset have been shown a poor performance when applied to real-world dataset. Despite the fact labeled real-world datasets are almost rare, so unsupervised and semi-supervised methods are of interest of some researchers nowadays. Since the performance of the model using unsupervised and semi-supervised methods are unable to match the performance of supervised learning methods, research is limited and results are inconclusive. So there is a scope of more research in this area. Another important data related issue is that real-world data is normally highly class imbalanced, as usually there are more truthful reviews in comparison with deceptive ones. This could be addressed through data sampling and ensemble learning especially with boosting and majority voting techniques. Also, data can be analyzed using graph theory as nodes with vertices and fed into the machine learning model as features which can be a new dimension in this area.

References

[1] K. Adhav, P. S. Z. Gawali, and P. R. Murumkar, "Survey on Online Spam Review Detection Methods," vol. 5, no. 6, pp. 7875–7876, 2014.

[2] A. Bhowmick and S. M. Hazarika, "Machine Learning for E-mail Spam Filtering: Review,Techniques and Trends," Jun. 2016.

[3] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," Heliyon, vol. 5, no. 6, p. e01802, Jun. 2019.

[4] A. Rastogi and M. Mehrotra, "Opinion Spam Detection in Online Reviews," J. Inf. Knowl. Manag., vol. 16, no. 04, p. 1750036, Dec. 2017.

[5] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," J. Big Data, vol. 2, no. 1, p. 23, Dec. 2015.

[6] N. Jindal and B. Liu, "Analyzing and detecting review spam," in Proceedings - IEEE International Conference on Data Mining, ICDM, 2007, pp. 547–552.

[7] E. P. Lim, V. A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in International Conference on Information and Knowledge Management, Proceedings, 2010, pp. 939–948.

[8] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting burstiness in reviews for review spammer detection," in Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013, 2013, pp. 175–184.

[9] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review spam detection via temporal pattern discovery," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 823–831.

[10] C. L. Lai, K. Q. Xu, R. Y. K. Lau, Y. Li, and L. Jing, "Toward a language modeling approach for consumer review spam detection," in Proceedings - IEEE International Conference on E-Business Engineering, ICEBE 2010, 2010, pp. 1–8.

[11] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, vol. 1, pp. 309–319.

[12] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam," in NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference, 2013, pp. 497–501.

[13] H. Sun, A. Morales, and X. Yan, "Synthetic review spamming and defense," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, vol. Part F1288, pp. 1088–1096.

[14] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What yelp fake review filter might be doing?," in Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013, 2013, pp. 409–418.

Authors' Profiles

Naznin Sultana received her Master's degree in Computer Science and Engineering from Jahangirnagar University, Bangladesh and completed her B.Sc. degree from the same institution in Electronics and Computer Science. Recently she has been admitted in PhD in Informatics in Malaysia University of Science and Technology, Malaysia.

She is currently servicing as an Assistant Professor in the department of Computer Science and Engineering at Daffodil International University in Bangladesh. Previously she served as a faculty member in Computer Science and Engineering department at two other reputed universities in Bangladesh. Her research interest includes Data Mining, Machine Learning, Image Processing and IoT.

Ms. Sultana is a member of Bangladesh Computer Society.

Dr. Sellappan Palaniappan is currently Professor of IT, Dean of School of Science and Engineering, and Provost of Malaysia University of Science and Technology. He has a PhD in Interdisciplinary Information Science from University of Pittsburgh (USA), a Master in Computer Science from University of London (UK), and a Bachelor in Statistics from University of Malaya (Malaysia). His current research interests include Data Mining, Machine Learning, Health Informatics, Web Services, Block chain, Cyber security, Cloud Computing and IoT.

How to cite this paper: Naznin Sultana, Sellappan Palaniappan, " Deceptive Opinion Detection Using Machine Learning Techniques", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.12, No.1, pp. 1-7, 2020. DOI: 10.5815/ijieeb.2020.01.01

