2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) | 978-1-6654-1460-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICT4SD50815.2021.9396899

Impact of Sentiment Analysis in Fake Online Review Detection

Rakibul Hassan, Md. Rabiul Islam
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
Email: rakibul.hassan@ece.ruet.ac.bd, rabiul.cse@gmail.com

***Abstract*—Online business is one of the rapidly growing business sectors of current world. Now-a-days people purchase a lot of things from online shopping sites. Sales of online products are most often review driven. Thus, detecting deceptive reviews is getting more importance day by day. Sentiment analysis has great importance in fake review detection system. This paper introduces a sentiment analysis model that can separate positive and negative sentimental reviews efficiently. It shows an analysis of sentiment distribution for fake and truthful reviews. Also, the proposed sentiment model is used to find the impact of probabilistic sentiment score in fake online review detection using a hotel review dataset.**

***Index Terms*—deceptive online reviews, sentiment analysis, probabilistic sentiment score, logistic regression, TF-IDF, support vector machine.**

I. Introduction

People don't need to go outside to buy necessary things now. Thanks to online shops and delivery websites. Every product like foods, clothes, electronics and others can be purchased and delivered to customer's home-place through these e-commerce sites. Sales of online products hugely depends on other's opinion who already have purchased the product and used it. To gain some knowledge about a product, it is one of the best ways to watch the review section and see opinions and comments given by other users. Thus, review section has a great impact in decision making [1]. This factor also provides the opportunity to some groups of dishonest businessmen or companies to manipulate public opinion about their products. They intentionally post fake reviews for promotion of their products. Also, there are some cases where some companies are attacked by their competitors. So, researches are growing interest in the field of fake review detection.

Various models have been proposed to detect fake online reviews. Machine learning approaches to detect fake online reviews include: supervised classification models, semi-supervised classification models and unsupervised clustering based models. Various features have been proposed by the researchers to create a better classification system. These features can be categorized into review-content based features, metadata based features, graph connectivity based features and user characteristics based features [2]. As fake user can always change their identity, content based features are most popularly used to detect deceptive online reviews. These features include word frequency count, n-grams, term frequency and inverse document frequency (TF-IDF), Parts-Of-Speech tag, noun to verb ratio and others [2]. Sentiment score as a feature is also used by many researchers. Fakes reviews are posted either to promote the products or demote it. Hence the probabilistic sentiment score is always much higher or lower when the review is deceptive.

In this research paper, we have introduced a sentiment model that can extract positive or negative sentiment effectively and with high accuracy. We have used the sentiment model to show the characteristics of sentiment score for both fake and truthful reviews. Also we have analyzed the effect of using probabilistic sentiment score from this model to build a fake review detection classifier. We have shown the comparison of classification results with probabilistic sentiment score from our sentiment model and binary sentiment score from the dataset.

The following section includes the related works section and the proposed sentiment model section. Section IV shows the results and findings of this research. The last section V, contains the scopes, future works and conclusions about our findings.

II. Related Works

Deceptive or fake review detection has found its attention from the very beginning of this century. Jindal et al. [1] introduced fake review detection as a classification problem in 2008. They categorized fake reviews into deceptive and destructive reviews. They described some features to classify deceptive reviews. Following, Ott et al. [3] in 2011, generated a gold standard dataset for fake online detection and they used n-gram features for classification. They updated their dataset in 2013 and added more fake reviews with negative sentiment. With more data with positive and negative sentiment, the previous n-gram model's result was decreased in accuracy. Later sentiment score as a feature was used by many researchers.

Fontanarava et al. [2] analyzed different features for fake review classification with supervised learning. They have categorized these features into two categories. These are -

- review centric category
- reviewer centric category

The review centric category includes text, text statistics, sentiment evaluations and metadata subsets. In the first three subsets, n-gram, word count, capital word ratio, emotional

978-1-6654-1460-9/21/$31.00 ©2021 IEEE

word with respect to total word, ratio of first-person pronouns etc. are included. Metadata subset includes rating deviation, singleton rating and burst features. Burst features can be used to determine a reviewer spammer group. It is an analysis of sudden activity with high concentration. They also suggested some new features like density, rating deviation mean, rating deviation with respect to local mean and early time frame as review centric new features. Maximum content similarity, word number average, rating entropy, activity time of the user, date entropy, date variance etc. are suggested by them as reviewer centric new features.

Rout et al. [4] used parts-of-speech, LIWC and sentiment score as their features with semi-supervised learning for detecting deceptive online opinions. They achieved an accuracy of 84% using these features. Following them, Hassan et al. [5] used fix sentiment score that tells the review is either positive or negative with some other features and got 86% accuracy with supervised classification approach.

Many other researches also have been carried out on text reviews for sentiment classification. Chang et al. [6] proposed a model that can extract sentiment effectively from text reviews with hotel review dataset. Their proposed sentiment analysis model was heuristic based. They suggested various methodology for feature engineering. These methods include BOW model, TF-IDF model and Doc2Vec model. The BOW model is the simplest model and can perform well to extract features with short text. But it has limitations of not considering position and semantic meaning of individual words. The TF-IDF model suggests that, all words are not similarly important. Words like article, prepositions, numbers, conjunctions etc. have less importance. The TF-IDF model adds weight with each word to tell its significance. TF-IDF model performs will to extract important features for sentiment analysis. Another model named WE model uses word embedding that transform a single word into a long vector. With decision tree and random forest classifier and using the following features Chang et al. [6] classified the correct sentiment efficiently. The ROC curve of the model covers 0.89 area under the curve.

III. Proposed Work

A. Dataset description

The dataset used for our proposed sentiment analysis model as well as fake review detection model includes 1600 hotel reviews with positive and negative sentiment. The proportion of positive and negative sentimental reviews is equal. Also the dataset is labeled with deceptive and truthful tag. This dataset was developed by Ott. et al [7] and is being used in many effective researches. The hotel reviews of Chicago, United Sates was collected by using Amazon mechanical Turk(AMT) and from websites like TripAdvisor and Yelp.com. The dataset is totally balanced with half deceptive and half truthful reviews. It only contains the review and does not contain user information and rating scores.

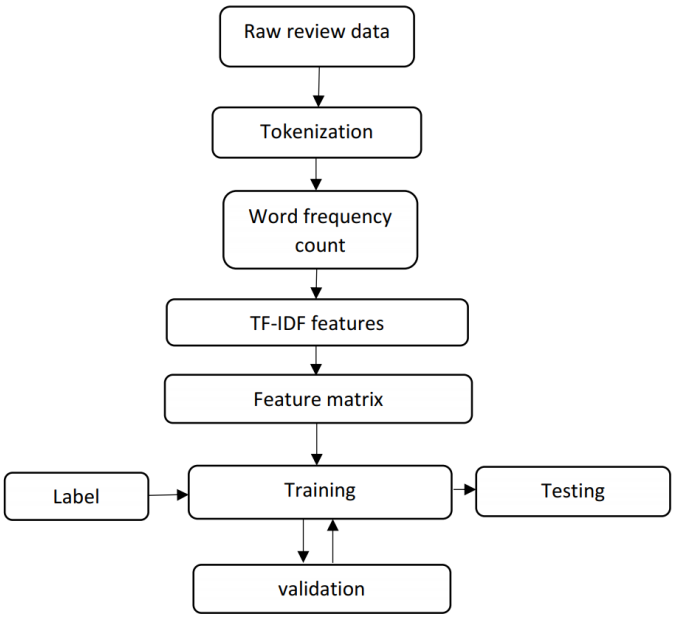

Fig. 1: Proposed sentiment model

B. Proposed methodology

a) Proposed sentiment model: In this research work we have developed a sentiment analysis model that can classify positive and negative sentiment effectively. We have used TF-IDF as a feature which is suggested as an important feature by Chang et al. [6]. Fig. 1 shows the structure of our proposed sentiment model. For constructing the model, first we have taken text reviews and generated tokens from the text. We have counted frequency of each words to calculate the term frequency (TF). Lets, $tf(t,d)$ denotes the term frequency of word t in the document d and $df(t)$ denotes frequency of t in the whole document. The inverse document frequency (IDF) can be calculated using the following equation.

$$idf(t) = log[\frac{n}{df(t)}] + 1 \tag{1}$$

here, n denotes the total number of reviews in the dataset and t denotes the $t'th$ word combining the TF and IDF we can calculate TF-IDF with the following equation-

$$tf - idf(t,d) = tf(t,d) * idf(t) \tag{2}$$

Using these TF-IDF features we have trained a classifier with labeled sentiments from the dataset. We have used the testing set to find the efficiency of our model.

b) Proposed fake review classification model: Our classification model uses the probabilistic sentiment score extracted from our proposed sentiment model with some other features. As TF-IDF is a good feature for text classification we have used it in the proposed fake online review classification model too. We have also used Empath features [8] which are linguistic features similar to LIWC and can be generated using open source tool.

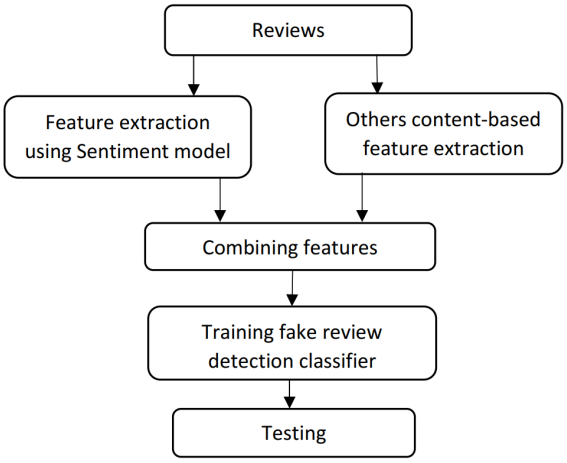

Fig. 2: Proposed fake review classification model

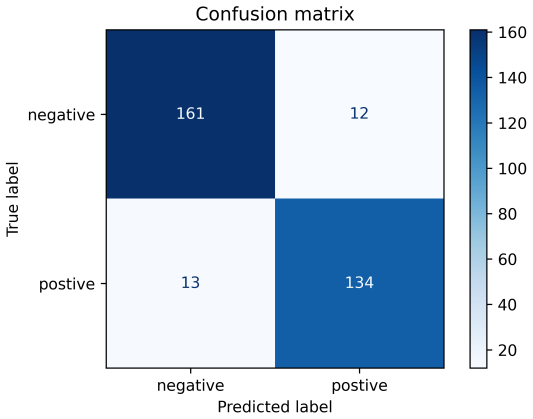

Fig. 4: Confusion Matrix for sentiment model

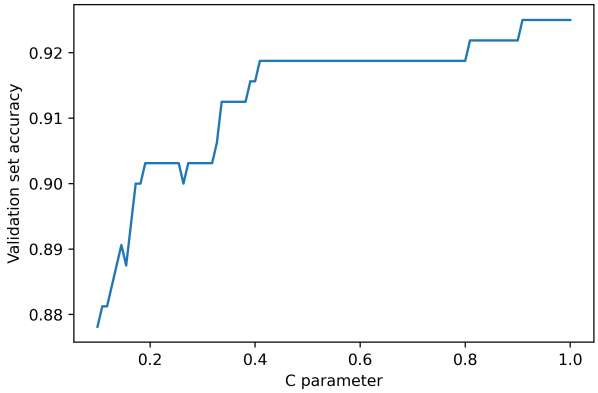

Fig. 3: C parameter tuning for logistic regression

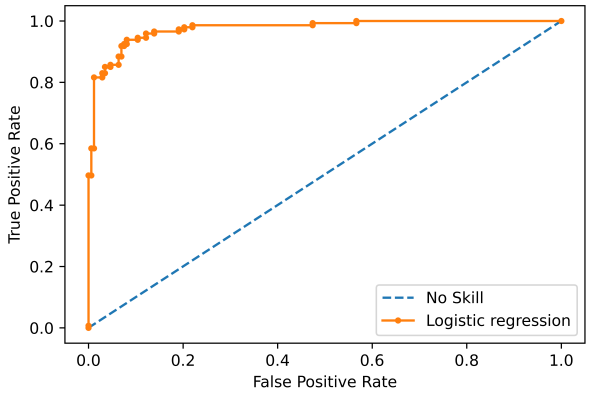

Fig. 5: ROC curve of proposed sentiment model

With several sets of other features, we have combined the probabilistic sentiment score to develop the feature matrix and used supervised machine learning classifier for classification and testing. We have also used fix valued sentiment score (0,1) instead of probabilistic sentiment score to show the impact of probabilistic sentiment scores generated by our sentiment model. The structure of our fake online review classification model is given in Fig. 2.

IV. PERFORMANCE ANALYSIS

A. Results

For building the sentiment model we have splitted the dataset into three set: train, validation and test set. The proportion of these sets were 60:20:20. We used logistic regression classifier for training. Using the validation set we have fine-tuned the C parameter of logistic regression classifier and set the value 1.0 for C. The validation set accuracy with different C parameters is given in Fig. 3.

With this set up we have got 92.18% test accuracy with 0.9178 precision, 0.9115 recall and 0.9146 F1 score. The confusion matrix from our testing is given in Fig. 4. Here we can see, among the 320 test examples, our sentiment classifier has classified 161 true negatives, 12 false positives, 13 false negatives and 134 true positive examples. The Receiver operating characteristics curve for our proposed sentiment model is given in Fig 5. The area under the curve (AUC) value is 0.975 which indicates a pretty good quality of our proposed model.

We have analyzed the probabilistic sentiment score for both fake and truthful reviews. Here we have a scale where 0 towards 1 indicates negative to positive. The average probability value for fake reviews with negative sentiment is 0.2504 and for truthful reviews with negative sentiment the average score is 0.2908. Again, The average probability value for fake reviews with positive sentiment is 0.7634 and for truthful reviews the average score is 0.7278. The probability distributions of sentiment score for both fake and truthful reviews is shown in Fig. 6

For fake review classification model, we have splitted the dataset into 80:20 train-test ratio. As features we have used TF-IDF, 194 categories from empath tool and binary sentiment score from dataset as well as probabilistic sentiment score from our sentiment model. We have used SVM classifier for

TABLE I: Comparative classification results

Works	Features	Classifier	Accuracy
Rout et al.[4]	Bigrams, Sentiment score, POS, LIWC	Logistic Regression	83.75%
Hassan et al.[5]	Word frequency count, sentiment score, Review Length	Naive Bayes	86.32%
Proposed work	Empath, sentiment score	Support Vector Machine	64.06%
	Empath, probabilistic sentiment score from sentiment model		66.56%
	TF-IDF, sentiment score		86.25%
	TF-IDF, probabilistic sentiment score from sentiment model		89.37%

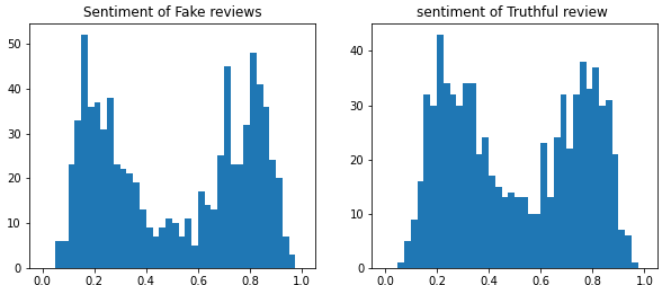

Fig. 6: Sentiment score distribution

fake review classification. With Empath and binary sentiment score we have achieved an accuracy of 64.06% where empath and probabilistic sentiment provides an accuracy of 66.56%. Also, TF-IDF with binary sentiment score provides 86.25% accuracy where TF-IDF with probabilistic sentiment score from our model provides 89.37% accuracy. The summary of our findings are shown in Table I.

B. Result analysis

In this research, we have developed a sentiment model which can classify positive and negative sentiment effectively. In the result section, we have shown different classification results with and without features extracted from our sentiment model. The results show that, our sentiment model helps the classifier to improve classification accuracy of fake online review detection system. The distribution of sentiment probability for fake and truthful reviews indicate that the fake reviews are either too positive or too negative.

V. CONCLUSION AND FUTURE WORK

We have developed a TF-IDF based sentiment classification model that can classify sentiment value with 92% accuracy. The probability distributions of fake and truthful reviews show deviation from each other. The fake reviews are of higher positive or negative polarity. The probabilistic sentiment score from our model helps the fake review detection system to improve it's performance. TF-IDF and sentiment features extracted from our model provides 89% accuracy in fake review detection. In this research, only text centric features are used for developing fake review detection model. In future, reviewer centric features can be combined with it. Also, metadata and rating behavior can be analyzed as feature sets. Also our research is based on English language reviews. It can be done in other languages. Different large datasets can also be used to find the effectiveness of this model.

REFERENCES

[1] N. Jindal and B. Liu, "Opinion spam and analysis," Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 219–230. ACM, New York, NY, USA (2008).

[2] J. Fontanarava, G. Pasi and M. Viviani, "Feature Analysis for Fake Review Detection through Supervised Classification," 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, 2017, pp. 658-666, doi: 10.1109/DSAA.2017.51.

[3] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "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 (ACL-HLT '11), vol. 1, pp. 309–319, Association for Computational Linguistics, Portland, Ore, USA, June 2011.

[4] J. K. Rout, A. Dalmia, and K. K. R. Choo, "Revisiting semi-supervised learning for online deceptive review detection," IEEE Access, Vol. 5, pp. 1319–1327, 2017.

[5] R. Hassan and M. R. Islam, "Detection of fake online reviews using semi-supervised and supervised learning," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-5, doi: 10.1109/ECACE.2019.8679186.

[6] V. Chang, L. Liu, Q. Xu, T. Li, C-H. Hsu, "An improved model for sentiment analysis on luxury hotel review," Expert Systems. 2020;e12580. https://doi.org/10.1111/exsy.12580.

[7] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam," in Proc. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol., 2013, pp. 497–501.

[8] E. Fast, B. Chen, M. Bernstein, . "Empath: Understanding Topic Signals in Large-Scale Text," May 2016, San Jose, CA, USA 2016 ACM, doi: 10.1145/2858036.2858535.