Empirical Study of Fake Reviews Detection of Online Reviews from E-Commerce Website

Phani K.Cheruku[1], Atul Kumar[2]

[1], [2]KIIT College of Engineering, Gurgaon, India
cheruku.phanikumari[at]gmail.com,atulkumar1508[at]gmail.com

Abstract: ***In this paper we present an empirical study of fake reviews detection algorithm .With the increase in internet usage, the demand for online servicing is growing rapidly, this leads to some threats like fake review. The users who used a product or service may give a genuine review, which makes it useful when other customers search for product/services. Whereas the online fake review may damage the customer sentiments and leads to negative impact on the product or services. Users' opinions are the main source of reviews for selected products or services. To get profit or popularity for a services or brands fake reviews are generally written to advance or downgrade the targeted items. Existing systems studied fake reviews but a strong detection technique is needed in this problem. The service sectors like restaurants, e-commerce product selling websites have significant impact on their business through reviews, their customers increase when the reviews are good and vice versa. This proposed system examines detecting fake reviews that have been evaluated in the Yelp restaurant domain.***

Keywords: Support Vector Machine, Term Frequency, Logistic Regression, Natural language processing, K-Nearest Neighbor, Decision Trees

1. Introduction

Word of mouth information is playing a major role in product sales, promoting services, which in turn customer reviews listed online in this computerized world. 52 percent of internet users use the internet for searching products online and 24 percent of users browse products for purchasing. Online reviews play a major role in e-commerce websites for product purchasing like electronic items, books, clothes, and other branded products. Similarly online movie review plays a major role in the entertainment industry. Hotels and restaurant reviews on online make tourism easier for customers these days.

These days consumers are checking the online reviews for making their purchase decision, thus reviews are important but the amount of data is high and sorting relevant information is too difficult. A real review should be written by the legitimate users who used the product or services and the content describing the review is also considered as an important factor. Promoting a product or services or demoting a product or services are based on fake reviews sometimes. Some of the business owners ask the employees to write the fake reviews to promote the services, in such cases, the reviews are fake and malicious, which may demote the legitimate services provided by the other service providers.

Yelp.com is one of the major restaurant chain information websites. This website uses a algorithm to find the illegitimate reviews. The algorithm is highly privatized and secured. In this proposed work, yelp.com reviews are used for studying purposes. The dataset is downloaded from yelp.com with labeled as real and fake. We used vectorizer to extract features and trained and tested and analyzed our work**.**

2. Literature Survey

2.1 Existing Technology

Fake review detection is done the taken dataset by applying feature extraction techniques

- CountVectorizer
- Ngram model
- TfidfVectorizer

Machine learning algorithm applied on the above extracted features

- Naïve Bayes
- Random Forest
- Logistic Regression
- SVM

2.2 Steps

The proposed application should be able to identify fake or real reviews. Feature extraction models used are n-gram, n-count and TF/IDF. We used classification models Naïve Bayes, Random forest, Logistic regression and SVM to predict the review type.

- Extract the feature using n-gram, n-count and TF/IDF
- Apply machine learning models Naïve Bayes, Random forest, Logistic regression and SVM
- Split train and test set
- On test set, apply machine learning algorithm Naïve Bayes, Random forest, Logistic regression and SVM
- Predict the review types
- Compare the machine learning algorithm's accuracy on each feature extraction model

3. Proposed Architecture

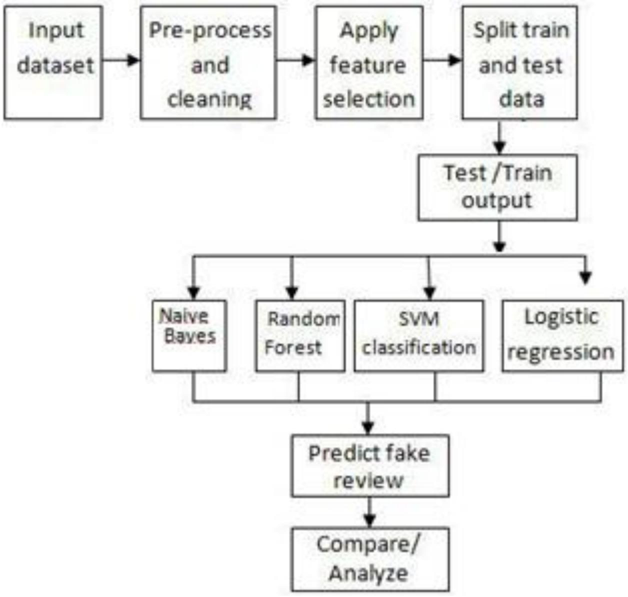

Figure 1.1: System Architecture of Fake Reviews Detection

4. Testing

Table 1.1: Test Cases of Proposed System

S. NO	Test Case ID	Test Descri-ption	Test Procedure	Test Input	Expected Result	Actual Result
1	T101	To check dataset loading	Load collected dataset	Execute fake review.py	Dataset should be loaded to execute	Error to load dataset
2	T102	To check correct dataset format	Load collected dataset	Execute fakereview.py	Dataset should be loaded to execute	Check the dataset field and column
3	T103	To check training	Start training dataset	Execute fakereview.py	Training should start and system learns data	Alert to user "Dataset is trained"
4	T104	To check prediction	Start prediction by test input	Execute fakereview.py	Test should start and output files generated	Alert to user "prediction completed"

5. Results

We have implemented Fake review detection from yelp dataset by applying three vectroization techniques namely CountVectorizer, Ngram model, TfidfVectorizer. The extracted features are trained and predicted using four machine learning algorithms namely Naïve Bayes, Random Forest, Logistic Regression, SVM. The proposed work is implemented in Python 3.6.4 with libraries scikit-learn, pandas, matplotlib and other mandatory libraries.

The following table shows the results arrive from our implementation model for N-gram feature extraction and prediction models. The following table shows the results arrive from our implementation model for N-gram feature extraction and prediction models.

Table1.1: Experimental Analysis of N-gram Model

Algorithm	Accuracy
Naïve Bayes	66.67
Random Forest	70.37
Logistic Regression	69.13
SVM	74.07

The following table shows the results arrive from our implementation model for N-count Vectorizer feature extraction and prediction models.

Table1.2: Experimental Analysis of N-count Model

Algorithm	Accuracy
Naïve Bayes	70.7
Random Forest	76.54
Logistic Regression	70.37
SVM	80.24

The following table shows the results arrive from our implementation model for TF-IDF feature extraction and prediction models.

Table 1.3: Experimental Analysis of TF-IDF model

Algorithm	Accuracy
Naïve Bayes	69.13
Random Forest	76.54
Logistic Regression	74.07
SVM	67.90

From the above results we can understand that Naïve Bayes model is giving good accuracy on prediction.

References

[1] Zheng, Y., C. Liu, et al. (2016), "Neural Autoregressive Collaborative Filtering for Implicit Feedback", Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. Boston, MA, USA, ACM: 2- 6.

[2] Zhao, Q., Y. Zhang, et al. (2017), "Multi-Product Utility Maximization for Economic Recommendation", Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. Cambridge, United Kingdom, ACM: 435-443.

[3] Li, D., G. Zhao, et al. (2015), "A Method of Purchase Prediction Based on User Behavior Log", 2015 IEEE International Conference on Data Mining Workshop (ICDMW).

[4] Jawaheer, G., Weller, P., & Kostkova, P. (2014), "Modeling User Preferences in Recommender Systems: A Classification Framework for Explicit and Implicit User Feedback", ACM.

[5] T. Jiang and A. Tuzhilin, "Segmenting customers from population to individuals: Does 1-to-1 keep your customers forever?", Knowledge and Data Engineering, IEEE Transactions on, vol. 18, no. 10, pp. 1297– 1311, 2006.

[6] Tang, B., H. He, et al. (2016), "A Bayesian Classification Approach Using Class-Specific Features for Text Categorization.", IEEE transactions on knowledge and data engineering 28(6): 1602-1606.

[7] Spirin, N.; Han, J. Survey on web spam detection: Principles and algorithms. ACM Sigkdd Explor. Newsl. 2011, 13, 50–64.

[8] Chakraborty, M.; Pal, S.; Pramanik, R.; Chowdary, C.R. Recent developments in social spam detection and combating techniques: A survey. Inf. Process. Manag. 2016, 52, 1053–1073.

[9] Peng, J.; Choo, K.K.; Ashman, H. User profiling in intrusion detection: A review. J. Netw. Comput. Appl. 2016, 72, 14–27.

[10] Keele, S. Guidelines for Performing Systematic Literature Reviews in Software Engineering; Ver. 2.3 EBSE Technical Report: Software Engineering Group, School of Computer Science and Mathematics Keele University, UK and Department of Computer Science University of Durham, UK: 2007; pp. 1–57.