A Fake Review Detection System Using NLP and Machine Learning Techniques

P. Aishwarya Sri,
Software Engineer,
Tata Consultancy Services,
Chennai, Tamilnadu, India.
patiaishwarya.sri@tcs.com

R. Vamshidhar Reddy,
Graduate Scholar, Business Analytics,
University of South Florida,
Tampa, Florida, United States.
vamshidharreddy@usf.com

ABSTRACT : ***Online surveys about the procurement of products have become the principle source of customers' belief. So as to pick up benefit or popularity, for the most part spam reviews are composed to advance or downgrade a couple of target items or administrations. This process is known as review spamming. In the previous years, a collection of different strategies have been proposed so as to explain the issue of spam reviews. The specialists classified the investigations dependent on how includes are removed from review datasets and various strategies and methods that are utilized to illuminate the review spam detection issue. What's more, this investigation has distinguished distinctive execution measurements that are generally used to assess the precision of the survey spam recognition models. This examination recognized that achievement variables of any review spam identification technique have interdependencies. Through deriving essential features from the content utilizing Natural Language Processing (NLP), it is probable to lead review spam detections utilizing different machine learning algorithms. Right now, there are various machine learning strategies that have been proposed to take care of the issue of review spam identification and the exhibition of various methodologies for characterization and discovery of survey spam. Research on techniques for Big Data is of passion, since there are a great many online reviews, with a lot all the more being produced every day. The essential objective of this paper is to describe how effective execution of the spam review detection model is and to accomplish better accuracy using the machine learning algorithms.***

Keywords - Spam review detection, Supervised learning, Unsupervised learning, NLP.

1. Introduction

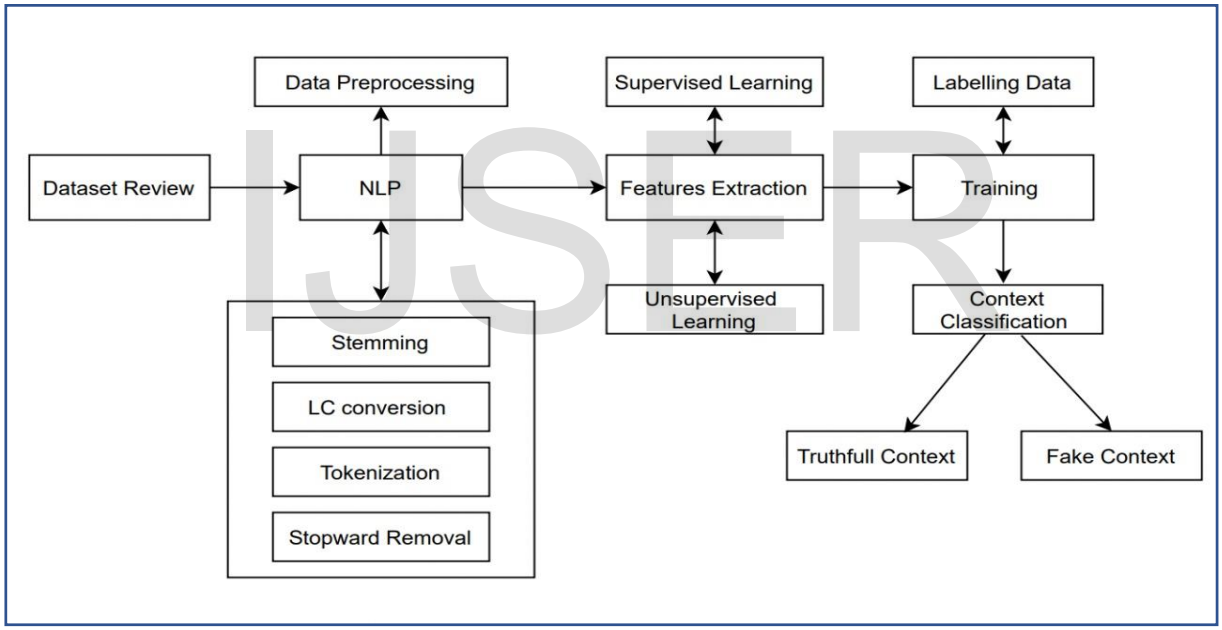
These days, websites have become the fundamental source for people to communicate. Human beings can effortlessly offer their perspectives about items and services by using online business sites, discussions, and web journals [1]. Most people read reviews about items and services before getting them. Everyone on the web is currently recognizing the significance of these online reviews for different clients and for sellers as well. Sellers are further planning additional marketing strategies depending on these reviews. For instance, if different clients purchase a particular model of a PC and compose reviews

in regards to issues related with its screen goals, at that time the producer may turn attentive and resolve this issue to expand consumer amusement.

One of the principle issues about review sharing sites is that spammers can effortlessly make publicity about the specific item by composing the spam reviews. These spam reviews may assume a key job in expanding the estimation of the item or service. For example, if a client needs to buy any item on the web, they generally go to the review segment to know about other purchasers' criticism. In this process, if that reviews are more positive, the client may purchase, else they would not purchase that item [2].

In most cases, spam review detection approaches comprise the following steps. The main fundamental step is the collection of the review dataset; since review datasets for the most part comprise of unstructured content and may contain noisy information, there is quite often a need to pre-process the datasets. The subsequent stage is to select an element designing methodology, for example, a linguistic n-gram or an individual spammer-based highlights approach. At long last, extraordinary review spam discovery systems will be evolved [3]. For example, machine learning and lexicon-based methods are applied to make sense of which reviews are spam.

2. Workflow



3. Datasets Review

The development or extraction of features from information is called feature engineering. Numerous examinations have utilized various sorts of feature engineering methods to remove the most widely recognized features or words in reviews. The most well-known component extraction strategy is the linguistic approach, and this procedure is applied by the bag of words approach [4]. In a bag of words technique, features for each review contain single words or little gathering of words found in a review content. Another feature engineering approach depends on a singular spammer's social qualities. Further, spammer features can be characterized into two kinds: review centric and reviewer centric features. Features that are developed utilizing data contained in a single review are called review centric features [5]. In reviewer centric features, this investigation not just considers all reviews composed by a creator on a similar stage like Amazon.com or Alibaba.com, also the additional data about the creator, for

example, review rating and time of posted review.

Accessibility of a dataset is the key beginning stage of any spam review detection. The key issue in the spam review detection issue is the accessibility of the labelled dataset. Scientists need to approach the labelled dataset to prepare a classifier with the goal that it might group an unknown review as spam or not-spam. The elective methodology is to utilize a falsely made review dataset by utilizing synthetic review spamming[6]. It is seen by the review of the writing that all review datasets are not openly accessible, and commonly scientists use crawlers to accumulate required information. It has been seen that a large portion of the analysts utilized Amazon.com web based business site datasets in their works [2], as it is the greatest web based business stage to have product reviews. Also, the specialists working in the spam review detection area utilize these datasets given by such sites. Through a couple datasets, for example, Amazon.com, Dianping, and Datatang [7,8,9], product reviews and in review datasets are freely accessible, anyway the issue of unlabeled information exists in their datasets also.

4. Preprocessing using NLP

a. Stemming: A stemming algorithm changes over various types of the word into a solitary identified form. For example, consider the words "works", "working", and "worked" as examples of the word work. Stemming must be applied to the review message before tokenizing it [10].

b. Tokenization: Here, words or gathering of words are utilized as characters. This linguistic technique is called uni-gram when a single word is chosen, bi-gram when two words are chosen, tri-gram when three words are chosen, etc. This strategy is called n-gram all in all. For instance, consider a review of "decent vehicle" and utilization of various n-gram methods on it. Unigram: ["good", "car"], Bi-gram: ["good car"], Uni + Bi-gram: ["car", "great", "great car"]. This work employed distinctive n-gram combinations on review information [11].

c. Word Embedding: Word embedding is the cutting edge method for characterizing words as vectors. The point of word embedding is to reclassify the high dimensional word features into low dimensional component vectors. An alternative way of representing words at a X and Y vector facilitates where related words, based on text of connections, are put nearer together. Word2Vec and GloVe are the most well-known models to change over content to vectors [12, 13].

d. Stopword Removal: Generally, the review content contains pointless words like "is", "the"," and", "a". These words are not useful in distinguishing spam reviews, accordingly, it is smarter to evacuate them before tokenizing to keep away from noise and irrelevant tokens. For example, take a review "This is a good vehicle". After expelling stop words and punctuation, the review shows up as a "good vehicle" [14].

5. Features Extraction

Machine learning is one of the most significant and outstanding ways for spam review detection and is commonly classified into supervised and unsupervised learning. Beneath, researchers talk about various machine learning techniques that have been proposed for spam review detection.

5.1 Supervised Learning

Supervised learning approaches utilized for spam review detection are commonly based on the classification strategies. Right now, two datasets are required: training data and test data. Training data is used to prepare the classifier and after that test data is used to assess the efficiency of a classifier. Procedures such as Support Vector Machine (SVM) and Naïve Bayes (NB) have an incredible accomplishment in review mining [15]. Researchers commonly start by collecting and dragging the dataset. The later stage is to prepare and pre-process the dataset accordingly. Once the dataset is arranged then the features are extracted from the dataset by utilizing the feature engineering approach. The next step is to prepare the classifier by utilizing training data. After all these steps, the efficiency of a classifier can be evaluated by utilizing test data.

A. Decision tree classifier: Decision tree (DT) classifiers give a various leveled breakdown of the training data space and are utilized to get familiar with the standards to recognize the realness of the review. A tree is shaped by utilizing various features and their qualities. Data gain is determined by utilizing a record of features. The feature that has greatest data gain is utilized as the root node of the decision tree. The inside nodes of the decision tree are named with remarkable features and these features have less information gain when compared with the root node. This technique is replayed until all reviews are segregated spam or on the other hand not-spam surveys [16].

B. Rule-based classifier: Rule-based (RB) classifiers utilize various standards to arrange spam or not-spam surveys. Rules might be applied to reviewer attributes, the substance of the review, or the product. A rule may be founded on text dimension, time to compose reviews, how frequently reviewers compose the reviews, length of the review, and how much of the time affectionate words like "bad" and, "good" are composed. The following four example rules explain the procedure of recognizing spam or not-spam review class [17].

Rule_1: In the event that a reviewer composes review 1 for item X and he again composes review 2 for item X in the next minute, at that point the subsequent review is with spam class.

Rule_2: If a reviewer composes review 1 for item X and he again composes review 2 for item X with a similar text dimension and style, at that point the subsequent review is with spam class.

Rule_3: If there are two reviews for a similar item and the length of the reviews is likewise the same, at that point the subsequent review will be considered as spam.

Rule_4: If a reviewer composes the review for an item with an excessive number of affectionate words, for example, "bad" and "good", the review is with spam class.

C. Probabilistic classifier: The probabilistic methodology is not quite the same as other methodologies such that specific changes between various reviews are revealed statistically instead of certain principles that are composed by a human or machine [18].

Bayesian Network: A Bayesian system shows the likelihood of the relationship among various nodes (features), and the feature is a component of a review that is being utilized to characterize the review. In addition, every node of the graphical model symbolizes a random variable and the edge represents the probability dependence between random variables. The connection between various edges is represented by Directed Acyclic Graphs. The probability of a node happening is the product of the probability that the random variable in the node happens given that the parents have happened. In the following condition, P(x1, ….xn) is the likelihood of any node xi and P(xi) is the probability of the parent [19].

$$P(x1, \dots \dots xn) = \pi\, P(xi \mid Pa(xi))$$

Naïve Bayes: This is a probabilistic classifier strategy dependent on Bayes' hypothesis. In addition, Naïve Bayes classifiers depend on the naïve assumptions that the features in a dataset are mutually independent. The following condition is the mathematical way of expressing Naïve Bayes classifier [20].

$$P(C|X) = P(X|C).P(C) / P(x)$$

D. Linear classifier: Linear classifiers use a linear combination of feature estimations of reviews and function well for the review classification issue, as it sets aside less effort to train when compared with a non-linear classifier. In linear classifiers, Support Vector Machine (SVM) characterization is most appropriate for the text data. This is a direct result of the sparse nature of the content where features are not identified with one another, yet they will associate to each other, and for the most part, these features are composed into independent classifications. Support Vector Machine strategy breaks down information and characterizes choice limits by having hyper-planes. In binary classification issues, the hyper-plane isolates the document vector in one class from another class, where the partition between hyper-planes is wanted to be kept as big as possible. Support Vector Machine optimisation method amplifies the predictive accuracy while consequently maintaining the distance from over-fitting of the training data. In addition, SVM projects the input information into the kernel space, and afterward it assembles a linear model. For a dataset (x1,y1), ... , (xn,yn), where y represents the class and x is the attribute which is associated with class y. As a result, any hyper plane can be composed as w.x - b = 0, where w is the ordinary vector to the hyperplane. SVM works very well for the limited quantity of training data and gives better outcomes for good tokenizers [21].

5.2 Unsupervised Learning

Freely accessible review datasets with the labelled classes are rare. Thus, unsupervised learning techniques that don't require a dataset with the class label are typically utilized. Unsupervised learning strategies drive the structure by examining the relationship between data, this technique is known as clustering. Data in one group is not at all the same with the data in the other group.

A. Twice-clustering technique: Twice-clustering is utilized to improve the exactness and variance of the unsupervised learning. Twice-clustering works in the progression of steps. To start with, the standard dataset which is unchanged is separated by utilizing k-fold cross validation. Second, all the training data to the group is selected for the first time to structure a cluster subclass and afterward clustering is applied to every subclass to frame an example subset of each subclass. The example subset of every subclass might be introducing some biasness. Hence, to solve this issue it is seen that non-uniform random sampling is a great way to deal with the structure of an example subset for every subclass. After all these steps, a subset of every subclass is chosen to develop a training set to train an unsupervised learner [22].

B. K-means clustering: K-means clustering has appeared to work admirably for a huge level of data and its precision level is likewise highly examined with other clustering algorithms. The K-means clustering algorithm gathers the separated terms as indicated by their feature values into K number groups, and K is any positive number that is utilized to find the number of clusters [23]. The K-means clustering algorithm performs the following steps.

1. Pick a number (K) of cluster centers.
2. Designate each item to its closest cluster center.
3. Move each cluster group to the mean of its designated items.
4. Perform the 2 and 3 until convergence is accomplished.

6. Training a classifier

Classifier or model is the algorithm that takes the input data and maps it to a specific type.There are many classification algorithms available now but it is not probable to confirm which one is better than the other. It relies upon the application and nature of the accessible data set. There are two fundamental steps in utilizing the classifier: training and classification. Training is the way toward taking content that is known to have a place with stated classes and making a classifier based on that known content. Classification is the way toward taking a classifier worked with such a training content set and running it on an obscure content to decide the class label for the obscure content. Training is an iterative procedure whereby you construct the most ideal classifier, and classification is a one-time process intended to run on obscure content.

7. Content classification

The text of a review is known as the content of the review. The content of each review is the principal thing to be considered in spam detection. Linguistic features, for example, word and POS n-grams for recognizing wicked practice (for instance, frauds and untruths) can be extracted from the content of a review [24]. Even Though linguistic features extracted from the text of a review are huge in spam detection, the approaches dependent on them are not adequately exhaustive to identify a wide range of fake reviews. An accomplished review spammer composes counterfeit reviews so easily that even a specialist in review spam identification may not be able to recognize it from a troop of honest reviews by basically perusing the content of the review. Depending on the type of review the content is classified into either truthful content or fake content.

8. Conclusion

Data is the most powerful weapon in digital technology. In modern times, online reviews are playing an important role in purchasing various products due to which the problem of fake reviews arised. Here, Datasets are reviewed and the features are extracted by using Natural Language Processing (NLP) [25] techniques. The features are classified and labelled by using the supervised and unsupervised machine learning algorithms. The data sets are trained continuously by classification algorithms and then tested on obscure text which is repeated until a unique classifier is developed. Finally, by using the ideal classifier the falsy reviews are identified and removed.

9. References

[1] R. Mohawesh et al., “Fake Reviews Detection: A Survey,” in IEEE Access, vol. 9, pp. 65771-65802, 2021, doi: 10.1109/ACCESS.2021.3075573.
[2] S. Tadelis, “The economics of reputation and feedback systems in ecommerce marketplaces,” IEEE Internet Computing, vol. 20, no. 1, pp. 12–19, 2016.
[3] S. Shojaee et al., “Detecting deceptive reviews using lexical and syntactic features.” 2013.
[4] C. C. Aggarwal, “Opinion mining and sentiment analysis,” in Machine Learning for Text. Springer, 2018, pp. 413–434.
[5] Crawford, M., Khoshgoftaar, T.M., Prusa, J.D. et al. Survey of review spam detection using machine learning techniques. *Journal of Big Data* 2, 23 (2015).
[6] N. Jindal and B. Liu, “Review spam detection,” in Proceedings of the 16th International Conference on World Wide Web, ser. WWW ’07, 2007.
[7] Julian McAuley Datasets available at: https://cseweb.ucsd.edu/~jmcauley/
[8] Dianping Datasets available at: http://liu.cs.uic.edu/download/dianping
[9] Datatang Datasets available at: http://www.datatang.com

[10] J. Plisson, N. Lavrac, D. Mladenic´ et al., "A rule based approach to word lemmatization," 2004.
[11] J. J. Webster and C. Kit, "Tokenization as the initial phase in nlp," in Proceedings of the 14th conference on Computational linguisticsVolume 4. Association for Computational Linguistics, 1992, pp. 1106–1110.
[12] Word to vector tool word2vec at: https://code.google.com/archive/p/word2vec/
[13] Vectors for Word Representation GloVe at: https://nlp.stanford.edu/projects/glove/
[14] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization," in Neural Networks, 2003. Proceedings of the International Joint Conference on, vol. 3. IEEE, 2003, pp. 1661–1666.
[15] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features." 1998.
[16] Patel, Harsh & Prajapati, Purvi. (2018). Study and Analysis of Decision Tree Based Classification Algorithms. International Journal of Computer Sciences and Engineering. 6. 74-78. 10.26438/ijcse/v6i10.7478.
[17] M. Almutairi, F. Stahl and M. Bramer, "A Rule-Based Classifier with Accurate and Fast Rule Term Induction for Continuous Attributes," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, pp. 413-420, doi: 10.1109/ICMLA.2018.00068.
[18] Large, J., Lines, J. & Bagnall, A. A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. Data Min Knowl Disc 33, 1674-1709 (2019). doi.org/10.1007/s10618-019-00638-y
[19] Bielza, Concha & Larranaga, Pedro. (2014). Bayesian networks in neuroscience: A survey. Frontiers in computational neuroscience. 8. 131. 10.3389/fncom.2014.00131.
[20] Wibawa, Aji & Kurniawan, Ahmad & Murti, Della & Adiperkasa, Risky Perdana & Putra, Sandika & Kurniawan, Sulton & Nugraha, Youngga. (2019). Naïve Bayes Classifier for Journal Quartile Classification. International Journal of Recent Contributions from Engineering, Science & IT (iJES). 7. 91. 10.3991/ijes.v7i2.10659.
[21] Mladenić, Dunja & Brank, Janez & Grobelnik, Marko & Milic-Frayling, Natasa. (2004). Feature selection using linear classifier weights: interaction with classification models. 234-241. 10.1145/1008992.1009034.
[22] Shi-fei Ding, Hui Li, Twice clustering based individual neural network generation method, Neurocomputing, Volume 157, 2015, Pages 264-272, ISSN 0925-2312, doi.org/10.1016/j.neucom.2015.01.007.
[23] S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 2010, pp. 63-67, doi: 10.1109/IITSI.2010.74.
[24] V. K. Vijayan, K. R. Bindu and L. Parameswaran, "A comprehensive study of text classification algorithms," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 1109-1113, doi: 10.1109/ICACCI.2017.8125990.
[25] G. G. Chowdhury, "Natural language processing," Annual review of information science and technology, vol. 37, no. 1, pp. 51–89, 2003.