International Journal of Engineering and Management Research e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume- 9, Issue- 2, (April 2019)
www.ijemr.net https://doi.org/10.31033/ijemr.9.2.12

Fake Product Review Monitoring & Removal and Sentiment Analysis of Genuine Reviews

Abhishek Punde[1], Sanchit Ramteke[2], Shailesh Shinde[3] and Shilpa Kolte[4]
[1]Student, Department of Computer Engineering, Ramrao Adik Institute of Technology, Navi Mumbai, INDIA
[2]Student, Department of Computer Engineering, Ramrao Adik Institute of Technology, Navi Mumbai, INDIA
[3]Student, Department of Computer Engineering, Ramrao Adik Institute of Technology, Navi Mumbai, INDIA
[4]Associate Professor, Department of Computer Engineering, Ramrao Adik Institute of Technology, Navi Mumbai, INDIA

[1]Corresponding Author: abhishekpunde1997@gmail.com

ABSTRACT

Any E-Commerce website gets bad reputation if they sell a product which has bad review, the user blames the e-Commerce website rather than manufacturers most of the times. In some review sites some great audits are included by the item organization individuals itself so as to make so as to deliver false positive item reviews. To eliminate these type of fake product review, we will create a system that finds out the fake reviews and eliminates all the fake reviews by using machine learning. We also remove the reviews that are flood by a marketing agency in order to boost up the ratings of a particular product .Finally Sentiment analysis is done for the genuine reviews to classify them into positive and negative. We will use Bag-of-words to label individual words according to their sentiment.

Keywords— Sentimental Analysis, Decision Tree Classifier, Native Bayes Classifier, Tokenization, Bag-of-words Tagging, Fake Reviews, Genuine Reviews

I. INTRODUCTION

As the vast majority of the general population require survey about an item before spending their cash on the item. So individuals go over different surveys in the site yet these audits are certified or counterfeit isn't identified by the client. In some review websites some good reviews are added by the product company people itself in order to make product famous this people belong to Social Media Optimization team. Client won't most likely find out whether the review is certifiable or fake. To find out fake review in the site this Product Review Monitoring and Removal and Sentimental Analysis of Genuine Reviews framework is presented. This framework will find out fake surveys made by social media optimization team by distinguishing the IP address. Client will login to the framework utilizing his client id and secret key and will see different items and will give survey about the item. To find out the audit is phony or certified, framework will find out the IP address of the client if the framework watch counterfeit survey send by a similar IP Address numerous multiple times it will illuminate the administrator to expel that survey from the framework. This framework utilizes data mining procedure.
This system helps the user to find out correct review of the product.

II. PRIOR APPROACH

Dictionary-based methodologies are valuable when a brilliant word reference is accessible that is important to the specialist or expert. One prominent kind of Dictionary is a sentiment dictionary which can be utilized to survey the sentiment of a given content via hunting down words that depict influence or conclusion. A portion of these word references are made by analyzing contrasting content based assessments of items in online gatherings to appraisals frameworks. Others are made by means of methodical perception of individuals composing who have been prepared to expound on various feelings.

III. ALGORITHM FOR SENTIMENTAL ANALYSIS

A. Native Bayes

Native Bayes' Theorem finds the probability of an occasion happening given the probability of another occasion that has just happened. Bayes hypothesis is expressed numerically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where A and B are events and P(B) not equals to 0.

Essentially, we are attempting to find probability of occasion A, given the occasion B is valid. Occasion B is likewise named as proof. P(A) is the priori of A (the earlier probability, for example probability of occasion before proof is seen). The proof is an attribute value of an unknown instance (here, it is occasion B). P(A—B) is a posteriori probability of B, i.e.

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume- 9, Issue- 2, (April 2019)
https://doi.org/10.31033/ijemr.9.2.12

probability of event after evidence is seen. Now, with regards to our dataset, we can apply Bayes' theorem in following way:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

where, y is class variable and X is a dependent feature vector (of size n) where:
$X = (x_1, x_2, x_3, ..., x_n)$

Native assumption

Now, its time to put a naive assumption to the Bayes' theorem, which is, independence among the features. So now, we split evidence into the independent parts.

Now, if any two events A and B are independent, then, P(A,B) = P(A)P(B) Hence, we reach to the result:

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

which can be expressed as:

$$P(y|x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1)P(x_2)...P(x_n)}$$

Now, as the denominator remains constant for a given input, we can remove that term:

$$P(y|x_1, ..., x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

Now, we need to create a classifier model. For this, we find the probability of given set of inputs for all possible values of the class variable y and pick up the output with maximum probability. This can be expressed mathematically as:

$$y = \text{argmax}_y P(y)\prod_{i=1}^{n} P(x_i|y)$$

So, finally, we are left with the task of calculating P(y) and P(xi — y).

Please note that P(y) is also called class probability and P(xi — y) is called conditional probability.

B. Decision Tree

Decision Trees are a kind of Supervised Machine Learning (that is you clarify what the information is and what the relating output is in the training data) where the information is constantly part as indicated by a specific parameter. The tree can be clarified by two substances, specifically decision nodes and leaves. The leaves are the final results. Also, the decision nodes are the place the information is split. There are two principle kinds of Decision Trees:

1. Classification trees (Yes/No types)
2. Regression trees (Continuous data types)

Here the decision or the outcome variable is Continuous.

C. Entropy

Decision tree is manufactured top-down from a root node and includes partitioning the information into subsets that contain occurrences with comparable qualities (homogeneous). ID3 calculation utilizes entropy to compute the homogeneity of an example. If the example is totally homogeneous the entropy is zero and if the example is a similarly isolated it has entropy of one. If we have more than 2 labels, we can generalize the Entropy formula as follows:

$$-\sum_i p_i \square log_2 p_i$$

where the p_i are the ratios of elements of each label in the set.

IV. OUR APPROACH

Sentimental Analysis using Supervised Learning

Step 1: Data Preprocessing : After the data has been selected, it needs to be preprocessed using the given steps:

1. Formatting the data to make it suitable for ML (structured format).
2. Cleaning the data to remove incomplete variables.
3. Sampling the data reduces the run time for algorithms and memory requirements.

Step 2: Tokenization: The process of breaking a stream of text up into phrases, words, symbols, or other meaningful elements called tokens. The goal of the tokenization is the exploration of the words in a sentence.

Step 3: Stop-word Elimination: The most common words that unlikely to help text mining such as prepositions, articles, and pro-nouns can be considered as stop words. Since every text document deals with these words which are not necessary for application of text mining. All these words are eliminated. A new list of stop words was created that eliminated only those words that did not contribute to opinion mining.

Step 4: Bag-of-words Model: The bag-of-words model is one of the simplest language models used in NLP. It makes an unigram model of the text by keeping track of the number of occurences of each word. This can later be used as a features for Text Classifiers. In this bag-of-words model you only take individual words into account and give each word a specific subjectivity score.

Step 5: Training the classifier: We are training the classifier using the Features Extracted using the Bag-of-Words Model. The Features of both the training and test dataset are com- pared. And this is giving to the classifier to give the predictions on the test data.

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume- 9, Issue- 2, (April 2019)
https://doi.org/10.31033/ijemr.9.2.12

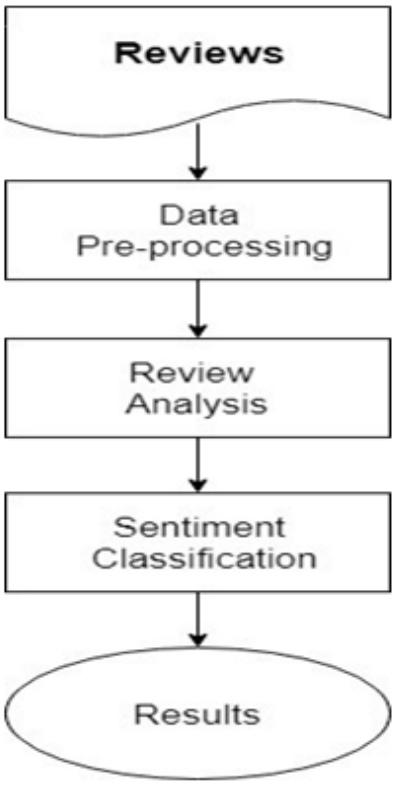

Fig. 1. **Sentimental Analysis Process**

Step 6: Sentimental Analysis: For sentimental analysis we are using the Decision tree classifier and Naive Bayes and comparing the results. We also see which classifier has the most accuracy.

A. SENTIMENTAL ANALYSIS

Sentiment analysis or emotion AI alludes to the utilization of natural language handling, content investigation, computational natural, and bio metrics to efficiently distinguish, separate, evaluate, and think about effective states and subjective information. Sentiment analysis is broadly connected to voice of the client materials, for example, reviews and overview reactions, on the web and web based life, and health care materials for applications that run from showcasing to clients administration to clinical drug. As a rule, sentiment analysis means to decide the frame of mind of a speaker, author, or other subject as for some point or the generally overall contextual polarity or emotional reaction to an archive, collaboration, or occasion. The attitude might be a judgment or assessment (see evaluation hypothesis), full of feeling state (in other words, the passionate condition of the creator or speaker), or the planned enthusiastic correspondence (in other words, the enthusiastic impact proposed by the creator or conversationalist) Fig 1. demonstrates the procedure for sentimental analysis.

B. SYSTEM DESIGN

It will be using Decision Tree Classifier and Naive Bayes Classifier for the purpose of sentimental analysis. We will be using the classifier to classify the words and label them into negative and positive. This System will help us to carry out fake product review removal and when the same person is giving reviews on the same product multiple times, our system will make sure that only the recent comment of the user is shown. If we find any suspicious IP and User-id adding review on the same product multiple times the admin will be notified. Fig 2. shows the flowchart for the proposed system.

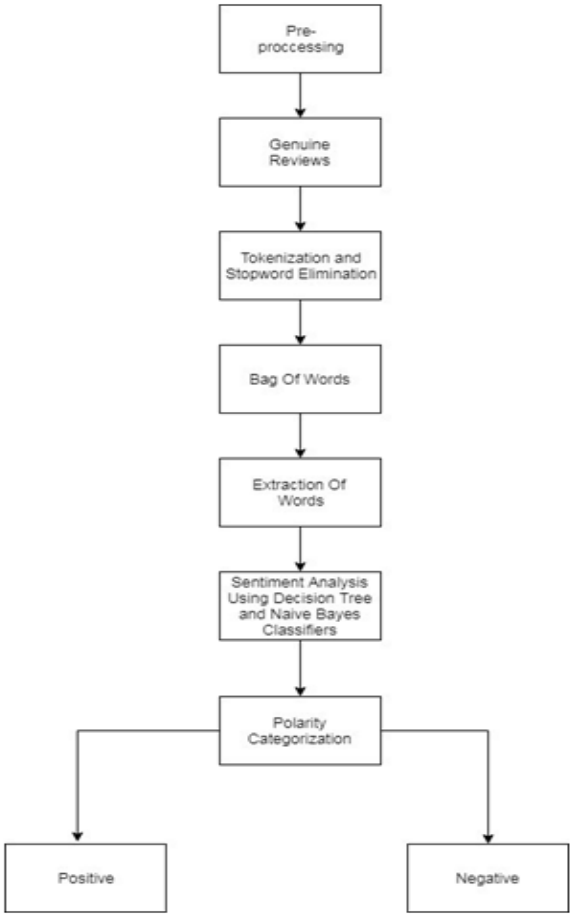

Fig2: **Flow Chart for the System**

C. RESULTS

Decision Tree Classification: The decision tree implementation obtained its best accuracy, after pruning the decision tree with a maximum depth of 10 on the Amazon Review dataset. The accuracy obtained is: 96.23%

Native Bayes Classification Model: For the Naive Bayes classification model, the priors are not needed to train the learning model. The Laplacian smoothing of 1.0 is applied

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume- 9, Issue- 2, (April 2019)
https://doi.org/10.31033/ijemr.9.2.12

to the Training dataset to estimate the value. We also used Gaussian NB but the Bernoulli NB gave a better accuracy of 95.90% .

V. CONCLUSION

In this Paper it is seen that sentiment analysis play vital role to make business decision about product/services. Sentiment Analysis not only encompasses concepts of text mining but also the concepts of information retrieval. Major challenges in Sentiment Analysis includes feature weighting which plays a crucial role for good classification. Also it is seen that soft computing techniques have not been extensively used in the literature. Without opinion life is like an empty vessel. The work can be further extended to emerging areas like Mobile learning and investigation with soft computing techniques like neural network.

Classification Model	Accuracy
Decision Tree Classifier	96.23%
Naïve Bayes Classifier	95.90%

Fig 3. **Comparison table between Decision Tree and Native Bayes Algorithm**

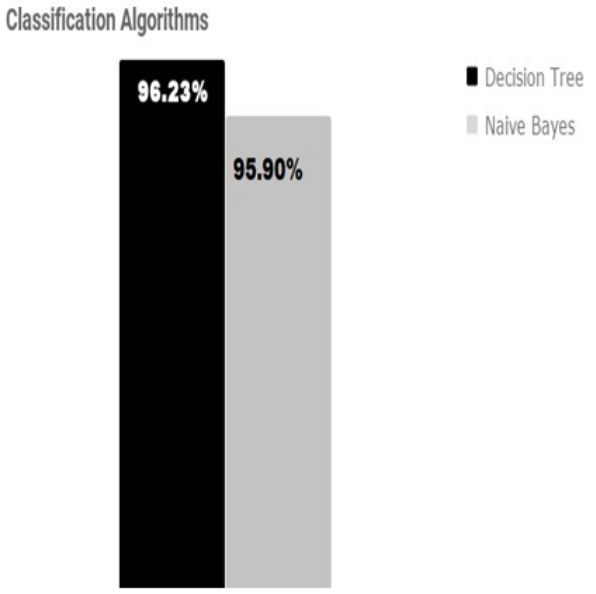

Fig. 4. **Comparison between Decision Tree and Native Bayes Algorithm**

REFERENCES

[1] Heydari, A., Ali Tavakoli, M., Salim, N., & Heydari, Z. (2015). Detection of review spam: A survey. *Expert Systems with Applications, 42*(7), 3634–3642.
[2] Lau, R. Y., Liao, S., Kwok, R. C. W., Xu, K., Xia, Y., & Li, Y. (2011). Text mining and probabilistic language modeling for online review spam detecting. *ACM Transactions on Management Information Systems, 2*(4), 1–30.
[3] Ramkumar, V., Rajasekar, S., & Swamynathan, S. (2010). Scoring products from reviews through application of fuzzy techniques. *Expert Systems with Applications, 37*(10), 6862–6867.
[4] Heydari, A., Tavakoli, M., & Salim, N. (2016). Detection of fake opinions using time series. *Expert Systems with Applications, 58*, 83-92.