iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.44368

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Detection of Fake Online Reviews using Semi-supervised and Supervised learning

Yashaswini D M[1], Prasanna H[2], Siddesh T[3], Dr. Sheela S V[4]
[1, 2, 3]*Student, Dept. of Information Science and Engineering, BMSCE, Bengaluru*
[4]*Professor, Dept. of Information Science and Engineering, BMSCE, Bengaluru*

Abstract: ***Nowadays, when somebody wants to make some decisions about a product or a service everyone goes with the reviews as it has become an essential part of decision making. When a customer wants to order a product on an e commerce website firstly everyone checks the review section in detail and further proceeds for decision making about the product. If the reviews posted were satisfactory for the customer he may order the product thus reviews become a reputed parameter for the businesses and companies and also a great source of information for the customers. Every customer thinks that the reviews he/she is seeing is authentic and any manipulation in that from any individuals or any rival companies which may lead to fake data which will be labeled as fake reviews. This type of attempt if not noticed may let us think about the gen-unity of the data. So these reviews are the most important parameter for the businesses and companies. There exist some groups or persons who make use of these reviews to forge customers for their own interest or damage their competitors reputation. In order to solve this problem we uses Machine learning techniques(Supervised and semi-supervised) to detect whether the given review is fake or not with high accuracy. Along with this objective we also focus on developing models which need less data to train.Since we can't always be able to get labeled data we use semi-supervised machine learning to make use of unlabeled data.It is understandable our model should be capable of giving results in reasonably less time. .In this paper we proposed many classification algorithm like Support Vector Machine algorithm (SVM) , Random Forest algorithm (RF) and deep neural network.***

Keywords: Support vector machine, deep neural network,semi-supervised, supervised, accuracy, e-commerce, product satisfaction

I. INTRODUCTION

It has become common for everyone to check online reviews before purchasing anything. This gives perfect opportunity for spammers to give fake reviews on their product to promote them self or to demote targeted products or companies. Estimates state that almost 4% of all online reviews are fake, which costed $152 billion. Since even a small company can hire online clients to give fake reviews easily detecting fake online reviews becomes an important issue to ensure the users don't get spammed easily. If that's not enough fake reviews can also be generated through bots, so it is very important to detect fake reviews.

Customers who purchase products from online firstly add similar products of different companies and make comparisons among them on which to buy. He/She mainly considers reviews as an important parameter while making decisions about buying it. Opportunistic individuals took advantage of this by defaming genuine products and promoting low quality products by providing good reviews for that through an individual or a set of groups. These are threats for customers, companies or businesses as their important parameters are being compromised while making decisions.

Online fake reviews are progressively growing up due to increase in e commerce and many of these instances are growing due to benefit companies from this. Due to recent pandemic people are forced to order through online, the number of users making online purchases sky-rocketed. So even if a small percentage of users gets affected because of fake review the cost will be enormous. As we all know this can happen again in the future we should be ready so detecting fake reviews not only help now but also it will be very helpful in the future. The methods based on machine learning techniques and deep neural network techniques were used to detect fake reviews that could mislead people. In this project we will overcome this problem.So Supervised and semi-supervised machine learning techniques can be used to identify fake reviews.

II. LITERATURE REVIEW

The research is done in detection of fake online reviews and this section explains the existing methodology work in classification of fake reviews or geninue reviews. N. KUMARAN explained using 2 models to solve the problems Authors as conducted a survey on several machine learning algorithms which helps to solve the problem of spam fake reviews detection. The classical machine learning methods are used detect speech, natural language and so on.

©IJRASET: All Rights are Reserved

Author as used python programming language because in python user can define his own variables and he can specify his own control structure for declaring them. Author is mainly evaluated on several machine learning algorithms by taking sample product and shopping data collected from yelp datasets.the experimental results clearly mention that random forest gives the better accuracy compared to other algorithms.

DEVARAPALLI SREEKAVYA proposes a novel convolution neural network model and bagging model is used to bag the neural network model the process goes like this first phase of tokenization then redundant words are removed and those words are created for candidate functionality.each feature is tested with dictionary and its frequency is counted and applied to the worlds numeric map.and it is also applied to functioning vector.finally the sentiment score is calculated and author as evaluated zero for negative review and 1 for positive review.Author has merged many features and tested which is not included in the previous work. So that the author as improved the precision of the semi supervised techniques.and maximum precision is given by supervised naive Bayes classifier.

CHAPALAMADUGU HARITHA CHOWDARY has used Hadoop data mining tool Generating datasets Processioning, Classification.author analyse the datasets based on accuracy given by naive Bayes multinational. Online review datasets accuracy around 94.8%. KAKKERA PRASANTHI Author as used supervised and semi supervised algorithms and expectation Maximization First tonkenization was done and redundant values must be removed. Sentiment score is calculated .author got the results as the precision of 81.34 % SVM gives the better accuracy rate.

AREMANDLA SAI PUJITHA So in this paper the author try to develop a model using semi-supervised technique to detect fake movie reviews. The approaches includes review content based which means this approach focus on the content of review and behavior of the user means checking the country, IP address, number of posts the user has posted and so on. There are three main techniques author used in this paper they are genre identification, detection of behavioral deception and text categorization. The reason for using these features are this allowed the author to reduce over fitting and to get high accuracy.

D.SAI KRISHNA author compares the performance of several experiments done on yelp datasets of a restaurants reviews with feature extracted and as well as without feature extracted from users behaviors. In both cases, the author compares the performance of several classifiers such as KNN, Naive Bayes, SVM, Logistic regression and Random Forest.

Kona Venkata Sai Mounic For detection of fake online reviews using supervised learning we are making use of some classification approaches which is Random Forest Classifier which will be used to improve the accuracy of our classifiers. Random Forest falls under supervised learning which uses an ensemble learning approach. Regression and classification both can be done using Random Forest. Random Forest makes use of many algorithms of the same type for its model which is multiple decision trees hence it is called Random Forest. It firstly collects random data points from the datasets. After random sampling it assigns a decision tree to it and the same will be applied to the number of decision trees being generated. And all trees will start training and produce their output. The right output will be taken into consideration by taking the majority of all those produced outputs. The datasets contains 2000 reviews which are in text format given by the user to a restaurant out of which 1000 are genuine and 1000 are fake reviews. Out of the datasets 80% is for training and 20% is for testing purposes which is a standard practice. After the training we need to check the accuracy and precision of the model. If it is found to be satisfactory then it is fine otherwise we need to retire the Random Forest model. Random forest classifier is used to classify the reviews. The datasets has three features: one review which is written by the user, polarity and the other one is fake or genuine.

Table 1.Classifiers and Results

Authors	Classifiers	Dataset	Results
N. KUMARAN[1]	Support vector machine Naïve Bayes Classifier	Hotel reviews datasets	When dataset is labelled well N aïve Bayes produces maximum precision when it is not vailable SemiSupervised learni ng works well.
DEVARAPALLI SREEKAVYA[2]	Decision Trees, NaiveBay e's	Amazon reviews datasets	decision tree accuracy is 88%, Naive Baye's accuracy is 77%

CHAPALAMADUGU HARITHA CHOWDARY[3]	NaïveBayes, Support Vector Machine and Decision Tree	restaurant reviews datasets	Naive Baye's accuracy is 90.31%. SVM accuracy is 83.75 % and Decision Tree algorithm accuracy is 66.56 %
KAKKERA PRASANTHI[4]	XGBoost, AdaBoost, and Gradient Boosting Machine for the classification of review	flipkart reviews datasets	Total review count = 67016 Fake = 8301 Genuine = 58715 Fake reviews % = 12.38 % Total reviewers = 34555 adaboost accouracy is 85%
AREMANDLA SAI PUJITHA[5]	roughset, decision tree, random forest and support vector machine.	News Dataset	The accuracy level of decision tree classifier is 97.7%.
D.SAI KRISHNA[6]	supervised learning, Random forests classifier	News Dataset	Using supervised technique Random Forest the author achieved highest accurace of 84%
Kona Venkata Sai Mounica[7]	SVM and naïve Bayes	Visiting places Database	frequency of occurred words with respect to their length is 175
D. Lalitha Bhaskari[8]	sentiment score shows a review's.	Hotel Dataset	Out of 627 product reviews, 10 reviews are found to be abusive, and hence, removed, and 48 are found to be spam
PILAKA ANUSHA [9]	Hadoop open source data mining tool	Movie review Dataset	Online review dataset accuracy around 94.968% and for twitter it around 82.695%
KAKI LEELA PRASAD[10]	Naïve Bayes Classifier	Movie review datasets	When dataset is labelled well Naïve Bayes produces maximum precision when it is not available SemiSupervised learning works well.
Dogo Rangsang[11]	Expectation Maximization using SVM, Naïve Bayes and Decision Trees.	The dataset consists of 3880 different reviews	Accuracy of SVM is 62.66%, Naïve Bayes is 62.79% and Decision Tree is 80.66%.
Seonget al.[12]	SVM	Hotel Dataset	frequency of occurred words with respect to their length is 175
ProposedSystem	SVM, RF,LSTM	USA news Dataset	SVM accuracy is 85%,RF accuracy is 100%,LSTM accuracy is 99.92%

Table 1 represents the description of the classifiers, datasets and results used from the research work. Based on research work most of them used the SVM and naive bayes classifier, for different types of datasets.

In the proposed system, model is implemented using the USA news dataset. Support vector machine, Random forest and LSTM algorithms are used for the USA news dataset which contains 2190 reviews with 2 different class. The accuracy achieved is 85% using support vector machine classifier, 100% using Random forest classifier, 99% using LSTM.

III. PROPOSED SYSTEM

Different data processing procedures that exist for detecting fake reviews. As per the study there are machine learning techniques which are already gift in our day life however provide less correct results. Thus, this system may be a smaller quantity correct and fewer effective. Also, it does not scale well for varied sorts of inputs.The proposed system is using the random forest and support vector machine and convolutional neural network for the classification of fake reviews detection. The gathering of dataset and preprocessing is incredibly vital because the choice of options ends up in the accuracy of the model.The dataset selected for training and testing phase plays a very important role.

A. System Architecture

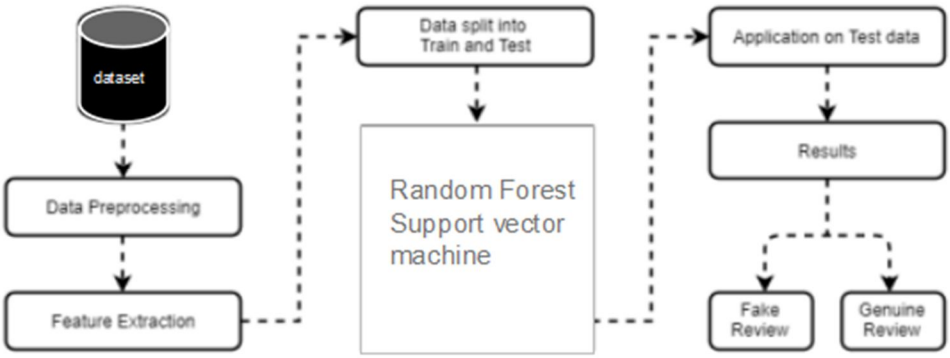


Figure 3.1. System architecture

The figure 1represents the proposed system architecture. The proposed system principally focuses on machine learning techniques. With the thought of the present challenges in fake reviews classification making an attempt to enhance accuracy mistreatment completely different classifiers. Using RF, SVM to get accurate results to classify the fake reviews. Both Random Forest RF and SVM classifiers offer correct results. The preprocessing of the data sets is incredibly vital in classification. The feature extraction and choice leads to the potency of the system. The american news dataset were used for training and testing. The planned model has its own benefits that it's simpler as compared to existing systems.

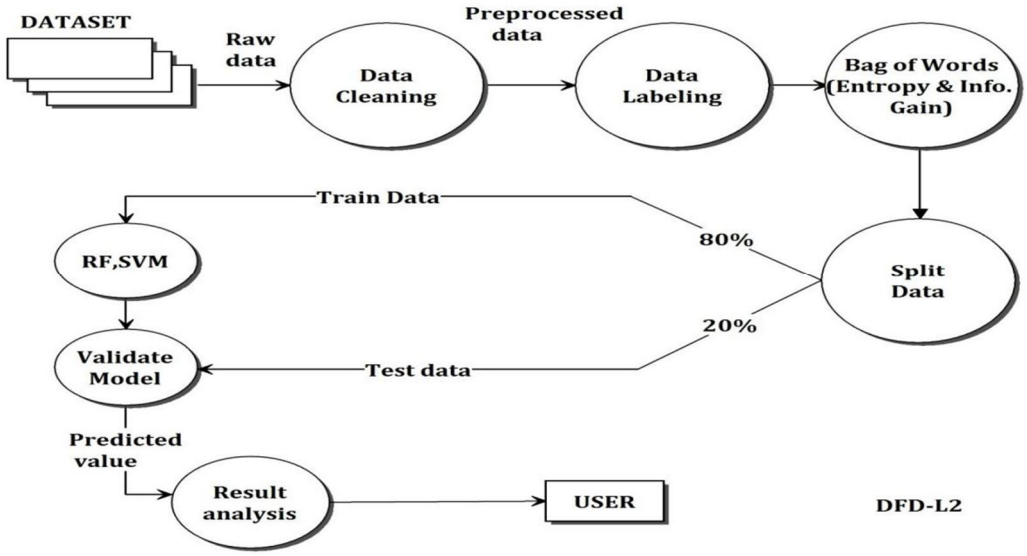


Figure 3.2. Detailed design of the proposed model

The figure 2. represents the detailed design of the proposed model.

The following area unit the steps to implement the given method

- Admin can add product to the system.
- The beginning preprocessing of knowledge takes place so useless columns is filtered out before the analysis method.
- The reviews containing express content and with swear words don't seem to be taken into thought and area unit faraway from the datasets.
- Sentiment score for every word is calculated then the words area unit extracted within the sort of wordbook that is termed as 'Bag of Words'
- Sentiment score of the review is calculated supported the size of -1 to +1.
- Spam removal is finished on the premise of their various options and analysis of product reviews.
- All the models area unit enforced and ending is explained and needed action is taken on analyzed reviews.

The following subsections explains the modules included in the proposed system. The descriptionof each module is as follows:

1) Data Collection

The first step includes the collection of data sets. The model is implemented using million song datasets. It consists of manyaudio files which are in wav file format. Each audio file is of 30 sec clips. The datasets consist of 10 different types of genres and each genre has 100 songs. The million song data sets contain genres like blues, classical, county, disco, hip hop, jazz, metal, pop, reggae and rock. 80% of the data is considered for training and remaining 20% for the testingphase.

2) Data Preprocessing

Data preprocessing may be a method of making ready the data and creating it appropriate for a machine learning model. it's the primary and crucial step whereas making a machine learning model. When making a machine learning project, it's not forever a case that we tend to come upon the clean and formatted knowledge. And whereas doing any operation with knowledge, it's necessary to wash it and place in a very formatted approach. therefore for this, we tend to use knowledge preprocessing task.

3) Model training

The model coaching is to be done once constructing the model. the info preprocessing part is vital because the feature choice improves accuracy of the model. The manual extraction of frequency and time domain options is finished. The classifiers random forest and support vector machine and CNN classifiers is supervised classification algorithmic program. It primarily depends on the amount of decision trees created. The additional is that the decision tree, additional is accuracy.

4) Random Forest

Random forest may be a supervised classification algorithmic program. RF trees also are referred to as as Random decision forests. it's used for classification, regression. It depends on the quantity of trees that exist. If there are a larger range of trees the a lot of is that the accuracy. during this the foundation node and have node split are done every randomly. The random forest creation and prediction are the 2 steps concerned during this algorithmic program. It operates by constructing a decision tree at training time and ultimately it labels the category..

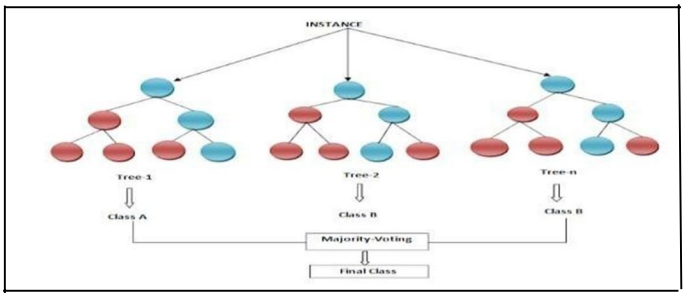


Figure 5. Random Forest Classifier

The figure 5.represents the operation of the RF classifier. The more is the construction of decision trees the higher is the accuracy for the classification. the biggest bulk of call trees is taken into account for any preprocessing.

5) Support Vector Machine

The support vector machine could be a supervised learning formula. The SVM could be a classifier that produces a bunch of hyperplanes in associate infinite-dimensional area. The SVM space units used for several functions like classification, regression. The creation of the most effective boundary observed as hyperplane is very very important in an exceedingly support vector machine. the advantages of victimization support vector machine space units, it'll modify high dimensional data sets, it's usually used to classify advanced biological detection of fake reviews data.

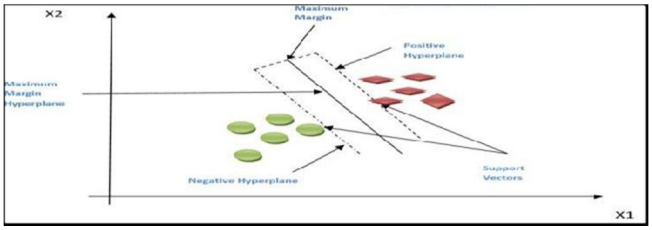


Figure 6. Support vector machine classifier

The figure 6. represents the support vector machine. The points that are nighest to the road are thought-about for the SVM. It consists of a hyperplane within which the positive and negative hyperplane is calculated for additional process of the information. If the separation between the categories is wider, svm tries to form a choice boundary.

6) Long short-term memory (LSTM)

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architectureused in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. For example, LSTM is applicable to tasks such as text classification, unsegmented, connected handwriting recognition, speech recognition and anomaly detection in network traffic or IDS's (intrusion detection systems), etc.

A common design consists of a cell (the memory a part of the LSTM unit) and 3 "regulators", typically referred to as gates, of the flow of data within the LSTM unit: an input gate, an output gate and a forget gate.Some variations of the LSTM unit don't have one or a lot of of those gates or even produce other gates. for instance, gated repeated units (GRUs) don't have associate degree output gate.

7) Model testing and evaluation

The last step once the model coaching is to check the model. At the testing part the testing is to be done on datasets. The model is evaluated exploitation trained knowledge and applied to test data set.Thus, within the testing part every part of the model is tested, and functionality of that part is checked. the most purpose of testing to envision whether or not every unit is functioning in success or failure state. using random forest, CNN and support vector machine classifiers, the training part is completed. The testing is completed using 20% of the dataset to predict the results.

The performance of models is evaluated using the following metrics.

Accuracy: Accuracy means the number of samples which are correctly classified to all possible samples.

IV. RESULTS

Implements Process:-

Step 1 :-In Command Prompt select the path of the file and Enter the python app.py

Step 2:-Click on the register button if your a new user and then submit it.

Step 3:-Click on the login button and enter the username and password and click on login.

Step 4:-Enter the text in the text box and click on the submit button.

Step 5:- Predict the Review is Fake or NOT Fake , by using RF, SVM, CNN-LSTM Algorithm's.

Step 6:- Finally Comparing three Algorithm's Accuracy and Predict the algorithm has highest Accuracy



Figure 5.1 Registering with valid parameters

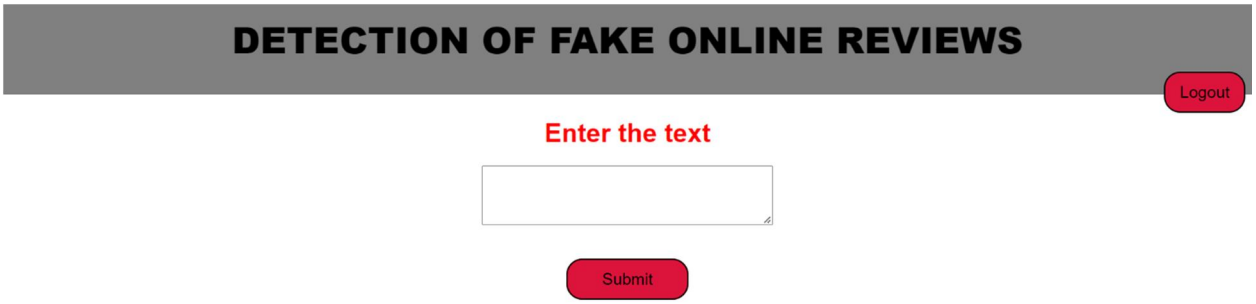


Figure 5.2 Results

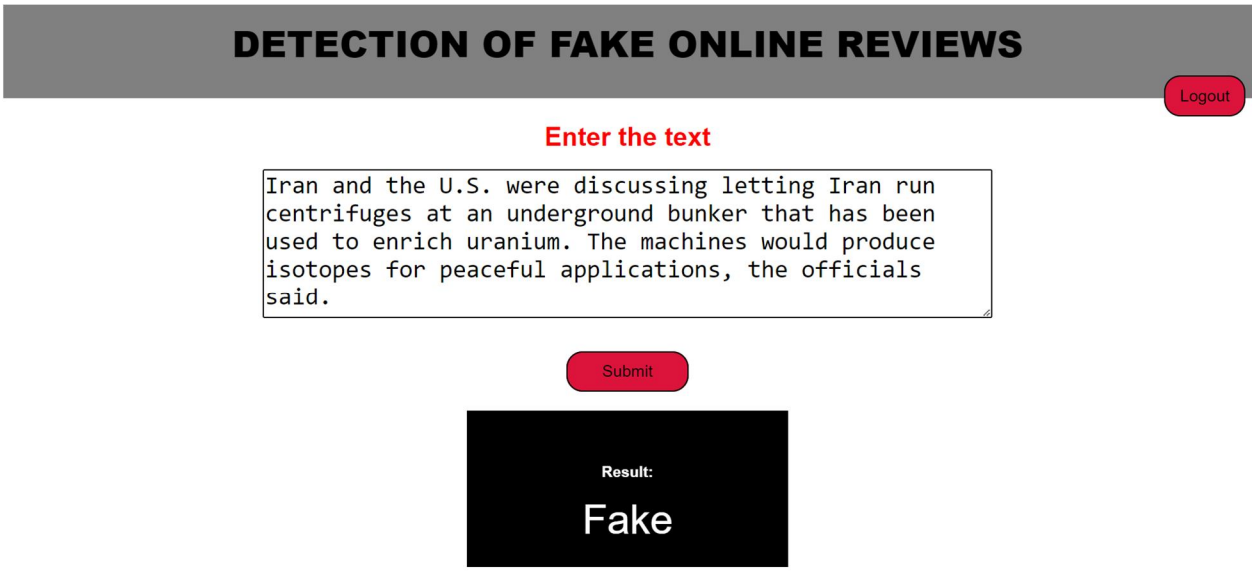


Figure 5.3 Results testing case 1

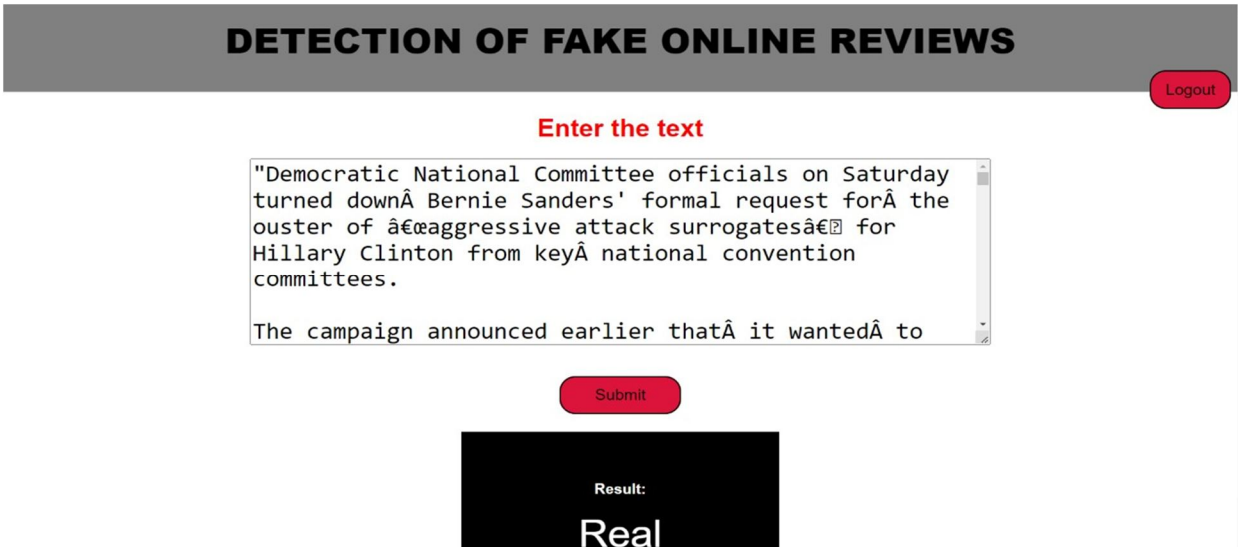


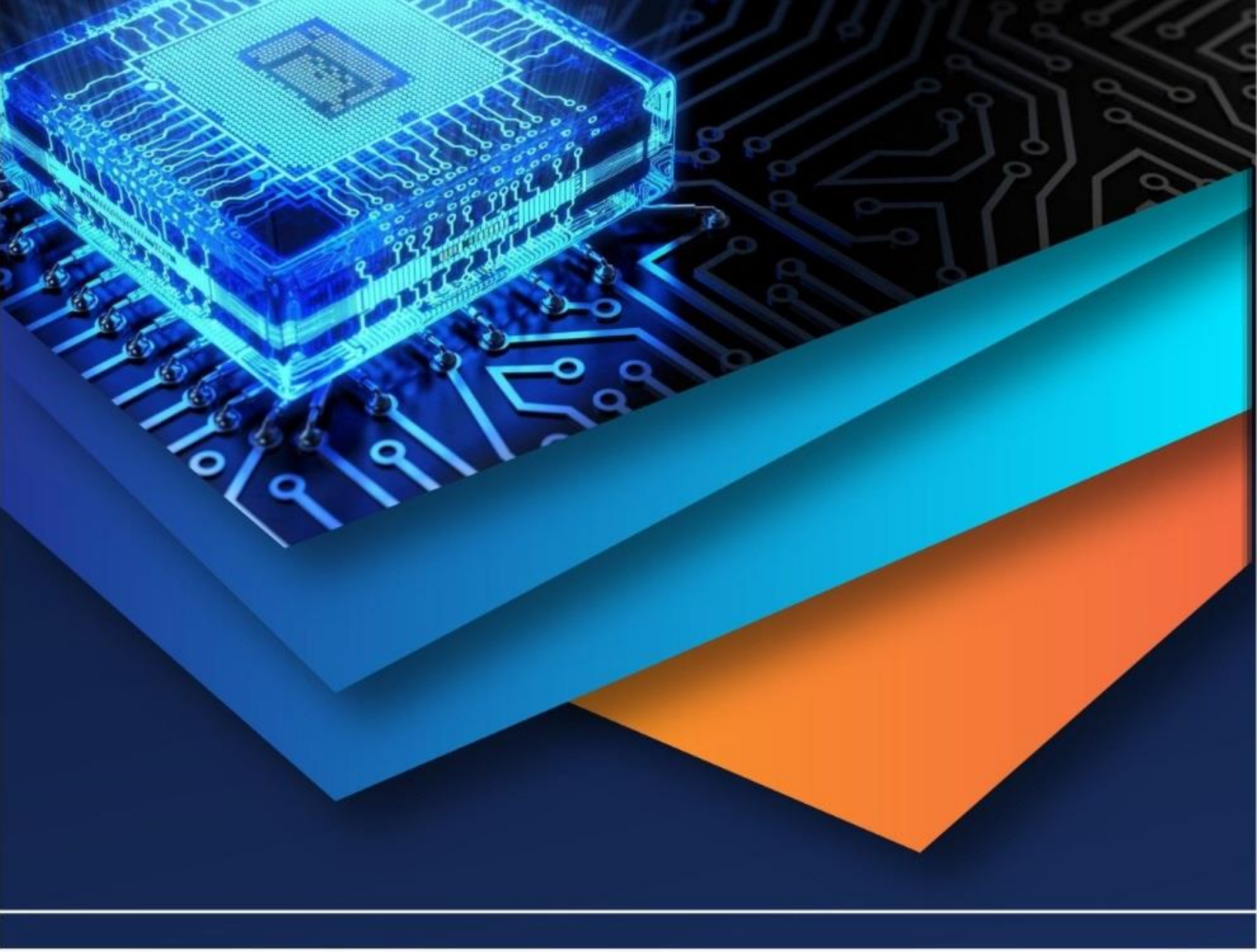
Figure 5.3 Results testing case 2

V. CONCLUSION AND FUTURE ENHANCEMENTS

We determined the random wooded area gives very great result. Hence it ensures our datasets is labelled properly as we apprehend semi-supervised model works nicely as on the identical time as dependable labelling isn't continuously available. In this task we've got were given in truth worked on client reviews. In future, client behaviours are mixed with texts to gather a much better model for classification. Advanced pre-processing equipment may be used for tokenization to make the datasets greater precise. In this paper we have used content-based approach for the detection of fake reviews, which means we focused on the content of the review i.e., textual part of review. But in the future one can use behaviour-based approach which means one can make use of country of the person that is making the review, IP address of the user, age, gender, ethnicity, number of reviews that the user give and so on. By including all this information one can make a better prediction to check the credibility of the review than checking the content of the review.

REFERENCES

[1] Chengai Sun, Qiaolin Du and Gang Tian, "Exploiting Product Related Review Features for Fake Review Detection," Mathematical Problems in Engineering, 2016.
[2] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: a survey", Expert Systems with Applications, vol. 42, no. 7, pp. 3634–3642, 2015.
[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), vol. 1, pp. 309–319, Association for Computational Linguistics, Portland, Ore, USA, June 2011.
[4] [4] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic Inquiry and Word Count: Liwc," vol. 71, 2001.
[5] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, Vol. 2, 2012.
[6] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a general rule for identifying deceptive opinion spam," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), 2014.
[7] E. P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), 2010.
[8] 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.
[9] J. Karimpour, A. A. Noroozi, and S. Alizadeh, "Web spam detection by learning from small labeled samples," International Journal of Computer Applications, vol. 50, no. 21, pp. 1–5, July 2012.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)