Volume 3, Issue , 2024
PP 305-314

International Journal of Futuristic Innovation in Arts, Humanities and Management (IJFIAHM)

DETECTION OF FAKE ONLINE NEWS USING MACHINE LEARNING IECHNIQUES

1st NandiniNarendraWairagade
bhavana.h2002@gmail.com
Department of Computer Science And Engineering
Shri Sai College of Engineering and Technology, Chandrapur ,India

2nd Prof LowleshYadav
lowlesh.yadav@gmail.com
Department of Computer Science And Engineering
Shri Sai College of Engineering and Technology, Chandrapur ,India

3rd ProfNeehalJiwane
neehaljiwane@gmail.com
Department of Computer Science And Engineering
Shri Sai College of Engineering and Technology, Chandrapur,India

ABSTRACT: Online reviews have great impact on today's business and commerce. Decision making for purchase of online products mostly depends on reviews given by the users. Hence, opportunistic individuals or groups try to manipulate product reviews for their own interests. This paper introduces some semi supervised and supervised text mining models to detect fake online reviews as well as compares the efficiency of both techniques on dataset containing hotel reviews.

Keywords: Fake online reviews, semi supervised learning, supervised learning.

INTRODUCTION: Technologies are changing rapidly. Old technologies are continuously being replaced by new and sophisticated ones. These new technologies are enabling people to have their work done efficiently. Such an evolution of technology is online marketplace. We can shop and make reservation using online websites. Almost, everyone of us checks out reviews before purchasing some products or services. Hence, online reviews have become a great source of reputation for the companies. Also, they have large impact on advertisement and promotion of products and services. With the spread of online marketplace, fake online reviews are becoming great matter of concern. People can make false reviews for promotion of their own products that harms the actual users. Also, competitive companies can try to damage each others reputation by providing fake negative reviews.

Opinion spamming is an immoral activity of posting fake reviews. The goal of opinion spamming is to misguide the review readers. Users involved in spamming activity are called "spammers". The task of a spammer is to build fake reputation (either good or bad) of a business by placing fake reviews. There exist some businesses who pay spammers to promote the company to attract new customers or to demote competent company of same type of business. A fake review either belong to positive or negative polarity. Review containing

praising statement about the product fall in “positive polarity”. And review containing loathing statements about the product fall in “negative polarity”. Increasing need for identifying fake reviews has captured the attention of researchers for solving the problem. Fake reviews not only mislead new customer for taking product purchasing decision but also affects business of good quality product. And due to false and misleading reviews on particular e-commerce site, users will avoid to visit that particular e-commerce site. It is concluded that identifying fake reviews will tackle three loses at one time.

Machine learning is a branch of Artificial intelligence, which is basically concerned with the development of algorithms .The data obtained is then processed by the algorithm is designed to identify complex relationships thought to be features of the underlying mechanism that generated the data, and employ these identified patterns to make predictions based on new data. It is the machine learning task of concluding a function from labeled training data .This function should predict the correct output value for any valid input object. The training data contains training examples or training values .

Supervised Learning: In supervised learning the training data consist of the input object and the Output object(Supervisory signal). It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. Basically supervised learning is when we teach or train the machine using data that is well-labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.

Fig No.1 Supervised Machine Learning

Unsupervised Learning: Unsupervised learning is used when the data is unlabeled or we have to find out the hidden structure. Since the examples given to the learner are unlabeled, there is no error to evaluate a potential solution. Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Semi-supervised learning is a type of machine learning that falls in between supervised and unsupervised learning. It is a method that uses a small amount of labeled data and a large amount of unlabeled data to train a model. The goal of semi-supervised learning is to learn a function that can accurately predict the output variable based on the input variables, similar to supervised learning. However, unlike supervised learning, the algorithm is trained on a dataset that contains both labeled and unlabeled data.

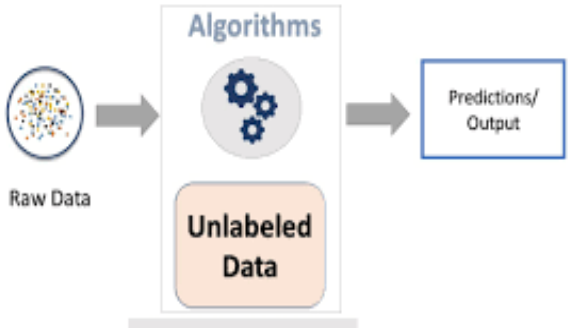

Fig No.2Unsupervised Machine Learning

Logistic regression

Logistic regression (LR) is a powerful and well-established method for supervised classification . LR helps in finding the probability that a new instance belongs to a certain class. Since it is a probability, the outcome lies between 0 and 1. Therefore, to use the LR as a binary classifier, a threshold needs to be assigned to differentiate two classes.

Support vector machine

Support vector machine (SVM) algorithm can classify both linear and non-linear data. It first maps each data item into an n-dimensional feature space where n is the number of features. It

then identifies the hyperplane that separates the data items into two classes while maximising the marginal distance for both classes and minimising the classification errors.

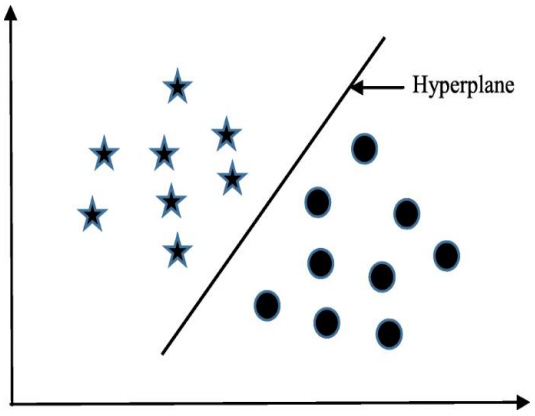

Fig No.3 Logistic Regression

Decision tree

Decision tree (DT) is one of the earliest and prominent machine learning algorithms. A decision tree models the decision logics i.e., tests and corresponds outcomes for classifying data items into a tree-like structure.

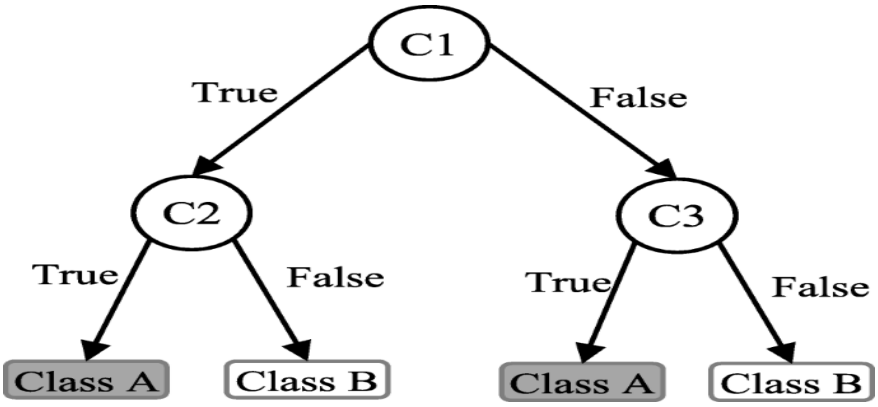

Fig No.5Decision Tree

Random forest

A random forest (RF) is an ensemble classifier and consisting of many DTs similar to the way a forest is a collection of many trees

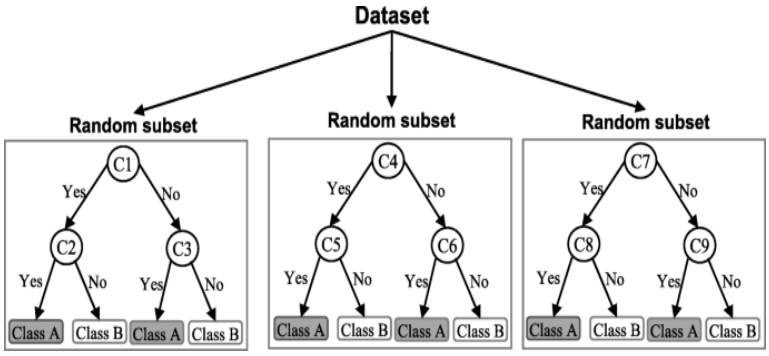

Fig No.6 Random Forest

An illustration of a Random forest which consists of three different decision trees. Each of those three decision trees was trained using a random subset of the training data

Naïve Bayes

Naïve Bayes (NB) is a classification technique based on the Bayes' theorem This theorem can describe the probability of an event based on the prior knowledge of conditions related to that event.

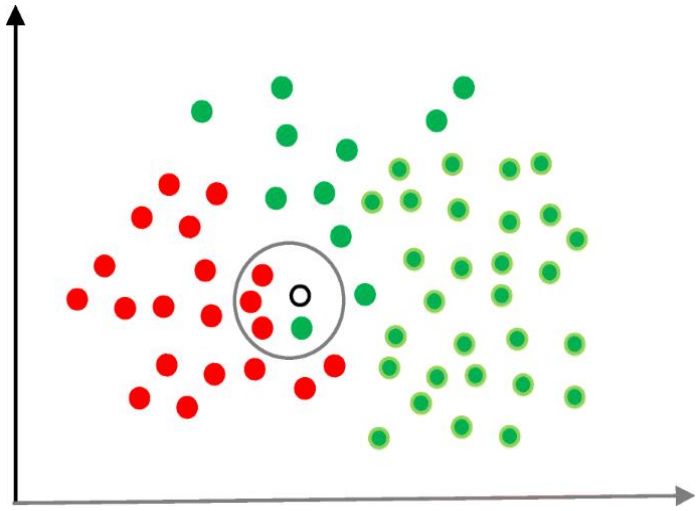

Fig No.7 Naïve Bayes

An illustration of the Naïve Bayes algorithm. The 'white' circle is the new sample instance which needs to be classified either to 'red' class or 'green' class

K-nearest neighbour

The K-nearest neighbour (KNN) algorithm is one of the simplest and earliest classification algorithms It can be thought a simpler version of an NB classifier. Unlike the NB technique, the KNN algorithm does not require to consider probability values.

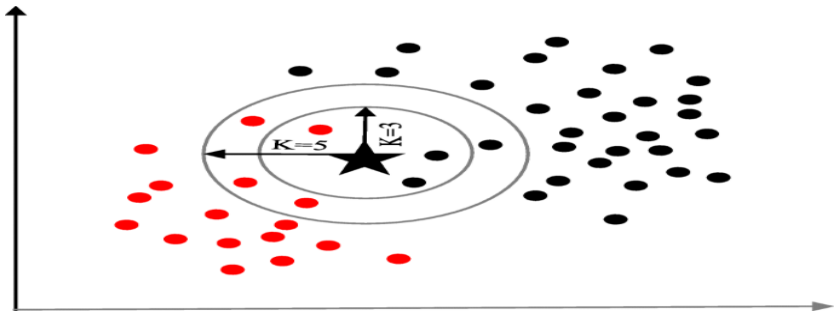

Fig No.8 KNN

IMPLEMENTATION:

In this paper, we make some classification approaches for detecting fake online reviews, some of which are semi-supervised, and others are supervised. For semi-supervised learning, we use Expectation-maximization algorithm. Statistical Naive Bayes classifier and Support Vector Machines (SVM) are used as classifiers in our research work to improve the performance of classification. We have mainly focused on the content of the review-based approaches. As feature we have used word frequency count, sentiment polarity and length of review

Data Pre-processing in Machine learning: Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model. Getting the dataset,

Importing libraries ,Importing datasets ,Finding Missing Data , Encoding Categorical Data , Splitting dataset into training and test set , Feature scaling Splitting the Dataset into the Training set and Test set In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model. Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So, we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

Dataset splitting: A dataset used for machine learning should be partitioned into three subsets training, test, and validation sets. Training set. A data scientist uses a training set to train a model and define its optimal parameters it has to learn from data. Test set. A test set is needed for an evaluation of the trained model and its capability for generalization. The latter means a model's ability to identify patterns in new unseen data after having been trained over a training data. It's crucial to use different subsets for training and testing to avoid model over fitting, which is the incapacity for generalization we mentioned above.

Model training: After a data scientist has preprocessed the collected data and split it into train and test can proceed with a model training. This process entails "feeding" the algorithm with training data. An algorithm will process data and output a model that is able to find a target value (attribute) in new data an answer you want to get with predictive analysis. The purpose of model training is to develop a model.

Processing: In many databases of real world contain conflicting and noise data. The reason is that data is often collected from numerous and heterogeneous sources. Inconsistency in data

results inaccurate outcomes in data mining process. Two types of preprocessing techniques are used for this research work: text and data preprocessing.

Text Preprocessing : Text preprocessing include data mining techniques used to transform unstructured text.

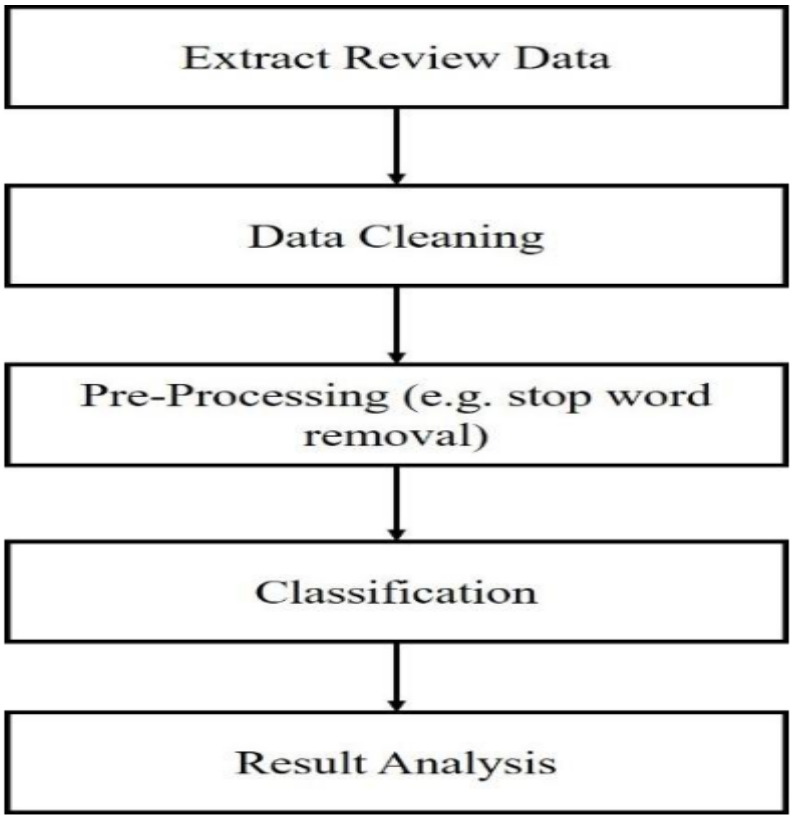

Fig No.9 Implementation

CONCLUSION

In conclusion, the detection of fake online reviews using ML techniques is a challenging but crucial task. Supervised and unsupervised learning algorithms, as well as hybrid approaches, have shown great potential in detecting fake reviews. Furthermore, recent studies have explored the use of deep learning techniques, which have shown promising results. However, the detection of fake reviews is an ongoing research area, and future studies should focus on developing more robust and accurate models to combat the rise of fake reviews on online review platforms.

FUTURE SCOPE:

In the future, a model that detects fake reviews as well as the reviewer which continuously spams the reviews using one or more accounts can be developed, and then accordingly a system to restrict or block such review accounts can be integrated with the model. Also as the dataset increases, the efficiency of the fake review detecting system also increases. So in the future, more datasets can be used to make the system more effective.

REFERENCES:

[1] N. KUMARAN, CHAPALAMADUGU HARITHA CHOWDARY, DEVARAPALLI SREEKAVYA, "DETECTION OF FAKE ONLINE REVIEWS USING SEMI-SUPERVISED AND SUPERVISED LEARNING", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 08 Issue: 04

[2] Lakkimsetty Suma, Smt.K.R.Rajeswari, Sri.V.Bhaskara Murthy, " Detection Of Fake Online Reviews Using Semi-Supervised And Supervised Learning ",Journal of Engineering Sciences Vol 13 Issue 07,2022, ISSN:0377-9254

[3] Kona Venkata Sai Mounica, D. Lalitha Bhaskari," Fake Online Reviews Detection Using SemiSupervisedAnd Supervised Learning" JETIR November 2020, Volume 7, Issue 11 (ISSN-2349-5162)

[4]Mr.M.THIRUNAVUKKARASU,P.KAVYA,MAHALAKSHMI.P.T,Mr.M.THIRUNAVUKKARASU ,"Fake Reviews Detection using Supervised Machine Learning", International Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 13, Issue 4, April 2023, pp. 250-254

[5] Venkatesh G S , Dr. N Gobi ,"Detection of fake online reviews using ML", International Journal of Advanced Research in Computer and Communication Engineering Vol. 12, Issue 4, April 2023

[6] Sharyu S Gadkari*, Prathamesh S Kore*, Mithila V Kulkarni*, Pratik M Zadbuke*, Prof. Puja Patil*5,"ONLINE FAKE REVIEW DETECTION USING MACHINE LEARNING" International Research Journal of Modernization in Engineering Technology and Science Volume:05/Issue:05/May-2023

[7] Ahmed M. Elmogy1 , Usman Tariq2 , Atef Ibrahim4," Fake Reviews Detection using Supervised Machine Learning"(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 1, 2021

[8] 1Pankaj Chaudhary, Abhimanyu Tyagi, Santosh Mishra," Fake Review Detection through Supervised Classification" ,IJCRT | International Journal of Creative Research Thoughts (IJCRT) , April 6-7, 2018 | ISSN: 2320-2882

[9] Rohinikhalkar1 , Murari Kumar Jha2 , Divyam Maru3 , Priyanshi Sharma4," Fake Reviews Detection using Supervised Machine Learning Algorithm" International Journal of Advances in Engineering and Management (IJAEM) Volume 4, Issue 7 July 2022

[10] Detection of Fake Online Reviews Using Machine Learning Techniques Dr. Sameena Banu1 , Asma Shaheen2 International Journal of Science and Research (IJSR) ISSN: 2319-7064

[11] Lowlesh Nandkishor Yadav,"Predictive Acknowledgement using TRE System to reduce cost and Bandwidth",IJRECE VOL.7 ISSUE1(JANUARY-MARCH 2019) pg no.275-278