DETECTION OF FAKE ONLINE REVIEWS BASED ON RANDOM FOREST USING MACHINE LEARNING

S.V.C Gupta, Professor,

A.Sowmya[2], P.Hema latha[3], Y.Anusha[4], CH.Manasa[5], Students

Dept of Computer Science And Engineering

Sri Vasavi Institute Of Engineering And Technology, Pedana, A.P, India

ABSTRACT:

Online reviews are very important in decision making of customer whether to purchase a product or service. These are main source of information getting from the past customer experience about the features of that service which we are going to purchase. This paper introduces some machine learning techniques like Random forest and Decision Tree for sentiment classification of reviews and to detect fake online reviews using the data set of a Hotel reviews. Sentiment Analysis has become most interesting in analysis of text. Using sentiment analysis we can separate negative and positive reviews as well. Online reviews have great impact on today's business and commerce. opinion reviews have an economical impact on the bottom line of businesses. Unsurprisingly, opportunistic individuals or groups have attempted to abuse or manipulate online opinion reviews (e.g., spam reviews) to make profits and so on.. Hence, opportunistic individuals or groups try to manipulate product reviews for their own interests.

Keywords: – Spam reviews, machine learning, Random forest Decision Tree algorithm.

INTRODUCTION:

A fake review is a misuse of the user reviewsystem by fake personalities. Fake reviews are also generated by bots. Fake reviews mislead customers to take decision on wrong product and the customer spends money on the product. There views can be either positive or Negative, to increase the promotion and sale or to bring down the competitive company products. Many people look at online reviews before making a decisionwhether it should be purchase or not. Many companies depend on several applications to detect Fake reviews using machine learning. In this paper we use Sentiment Analysis to formulate the data.The sentiment is usually formulated as a two-class classification problem,positive and negative. The basis of Sentiment Analysis is detecting the polarity of a give text ordocument. In this project we are using a set

DOI : 10.36893.DRSR.2021.V09I01N01.834-840

Polarity as negative or positive. In machine learning based techniques, there are many algorithms can be applied for the classification and prediction. Here we used Random Forest Classifier and Decision Tree for predicting the reviews. We detect fake positive, fake negative, True positive and True negative reviews. While used to detect fake reviews, supervised learning approaches suffer from several limitations. For example, unless one can be assured of the ‘‘quality’’ of the reviews used in the training dataset, we will have a garbage-in-garbage-out situation. In addition, the amount of labeled data points used to train the classifier can be difficult to obtain and update, given the dynamic nature of online reviews. In the authors highlighted that human are poor in labeling reviews as fake or genuine. This complicates the task of finding ground truth for given instances accurately. finally we compare the accuracy of each algorithm. Thus, in this paper, we use several Random forest learning approaches to improve the classification, to obtain better results. We then evaluate the proposed approach using a dataset comprising both positive and negative reviews.

LITERATURE SURVEY:

Deceptive online review detection is generally considered a classification problem and one popular approach is to use supervised text classification techniques. These techniques are robust if the training is performed using large datasets of labeled instances from both classes, deceptive opinions (positive instances) and truthful opinions (negative examples) explained that identification of deceptive online reviews is often performed using prior human knowledge, which increases the probability of mislabeled reviews due to the potential for subjectivity during the labeling process. It is also easier to collect a large amount of unlabeled reviews, in comparison to labeled reviews required in the training of supervised techniques. Thus, if we have a large number of unlabeled reviews, a viable approach is to use semi-supervised techniques. For example, Li et, used review and reviewer features to design a two-view semi-supervised method, by employing the framework of co-training algorithm to detect spam reviews.

However, the use of co-training in classification suffers from several drawbacks. For example, the manually labeled reviews used in co-training can be unreliable due to human involvement and subjectivity (reported only a 60% accuracy rate). The use of only positive and unlabeled data leads to poor performance in co-training algorithms . Such approach also does not consider the features of deep syntax and psychological linguistics of review text, which can help improve the effectiveness of deceptive review detection. The user cannot easily identify this kind of opinion spam. They have mined all 5-star truthful reviews for 20 most famous hotels in Chicago area from trip advisor and deceptive

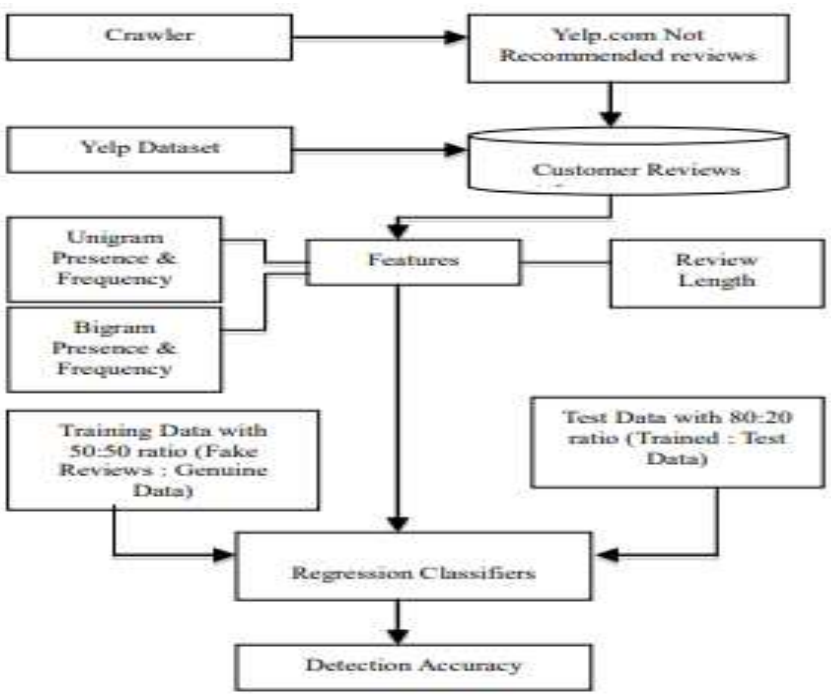
DOI : 10.36893.DRSR.2021.V09I01N01.834-840

opinions were gathered for the same hotels using amazon mechanical trunk (AMT). They first asked human judges to evaluate the review and then they have automated the task for the same set of reviews, and they found that automated classifiers outperform humans for each metric.

PROPOSED METHOD:

Data extracted from amazon dataset is used as the unlabeled data, labeled dataset is used for both training and testing purpose in this method. Preprocessing procedures includes- tokenization & lowercasing letters, removing stop words, removing punctuations, stemming etc. Active learning is a special case of semi supervised machine learning which can interactively request the user to determine the class of some unknown data points to achieve the desired results. Random Forest Random Forest algorithm is a supervised classification algorithm. Random Forest is the processes of finding the root node and splitting the feature nodes will run randomly. It is Unexpected accuracy among current algorithms.

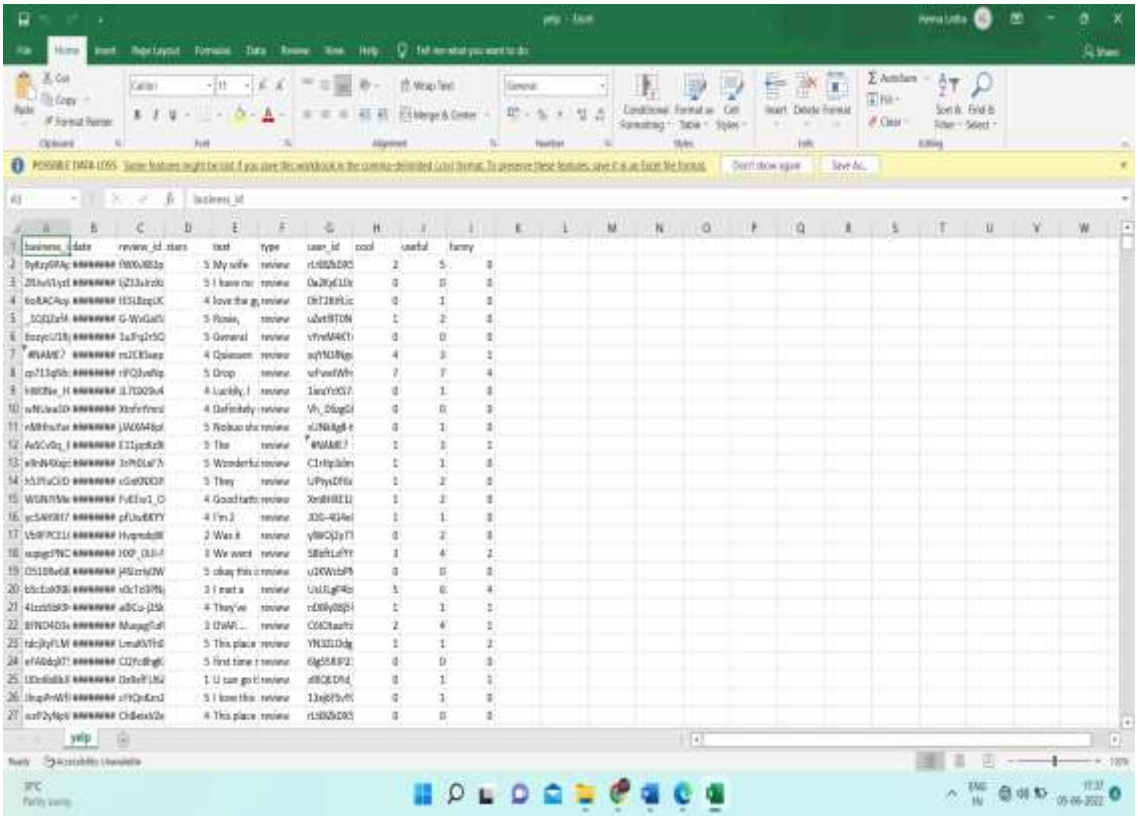
ARCHITECTURE:



DATASET:

DOI : 10.36893.DRSR.2021.V09I01N01.834-840

IMPLEMENTATION:



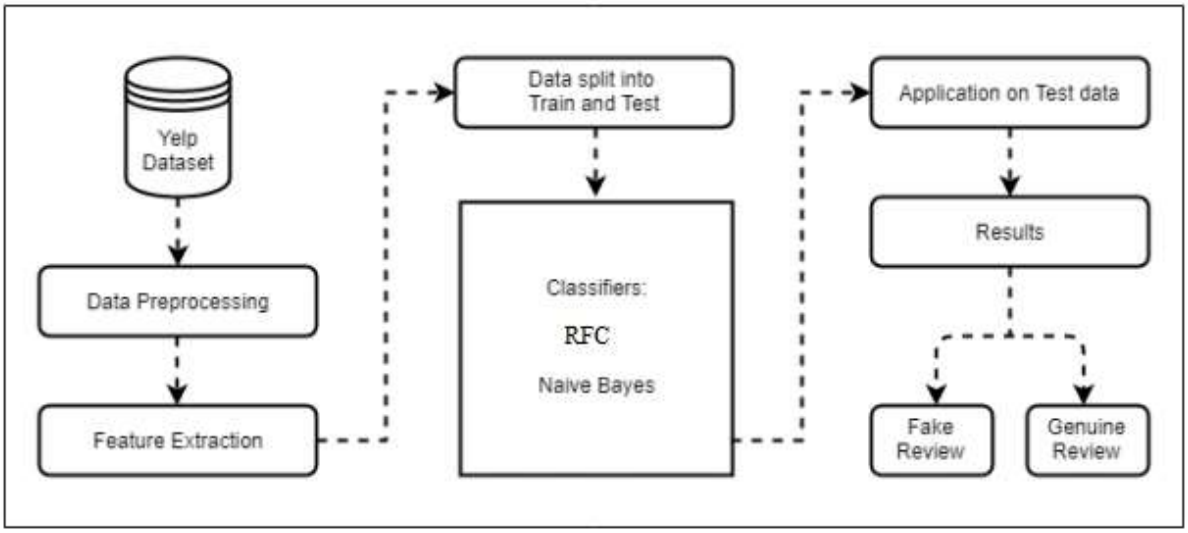
The Yelp Challenge Dataset includes hotel and restaurant data. Yelp's filter has been used as a reference for labelling reviews as fake or not. Yelp's filtering algorithm has evolved over the years to filter fake reviews . Also, this filter has been claimed to be highly accurate. we are expanding on the current literature by introducing ensemble techniques with various linguistic feature sets to classify reviews from multiple domains as true or fake. The ensemble techniques along with Linguistic Inquiry and Word Count (LIWC) feature set used in this research are the novelty of our proposed approach.

RANDOM FOREST ALGORITHM:

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. The given dataset and takes the average to improve

DOI : 10.36893.DRSR.2021.V09I01N01.834-840

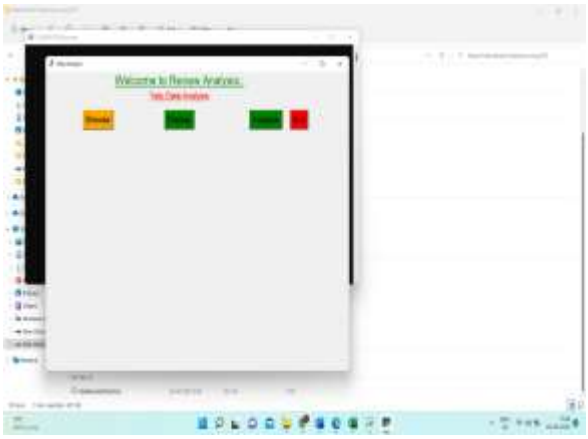
the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

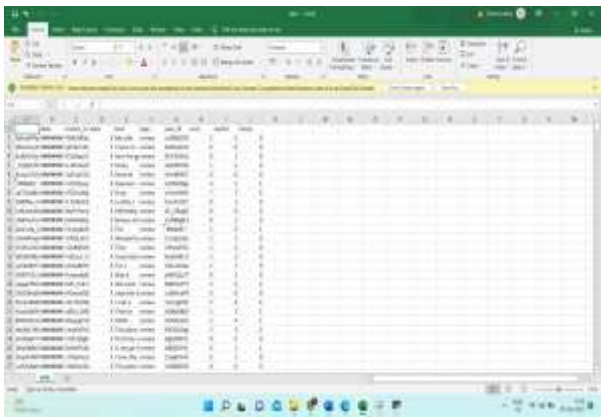


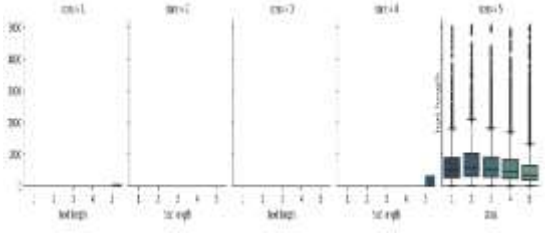
IMPLEMENTATION RESULTS:

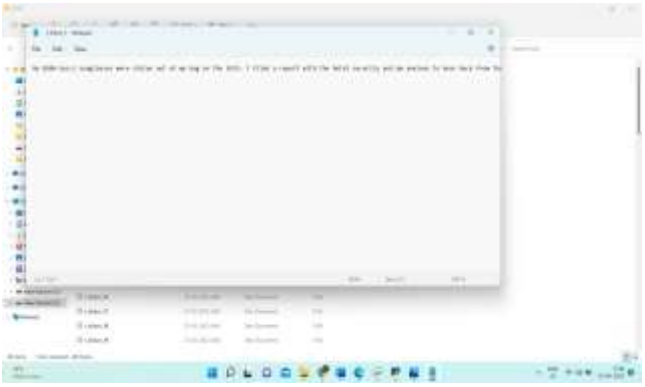
The results procured from each of the four methods are good, yet that doesn't show that the recommender framework is ready for real-life applications. It still need improvements. Predicted results show that the difference between the positive and negative class metrics indicates that the training data should be appropriately balanced using algorithms like random forest bayes classifier etc. Proper hyper parameter optimization is also required for classification algorithms to improve the accuracy of the model. In the recommendation framework, we simply just added the best-predicted result of each method. For better results and understanding, require a proper ensemble of different predicted results .

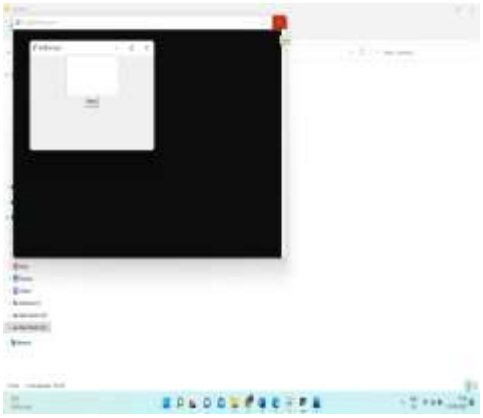
DOI : 10.36893.DRSR.2021.V09I01N01.834-840











DOI : 10.36893.DRSR.2021.V09I01N01.834-840

CONCLUSION:

Reviews are becoming an integral part of our daily lives; whether go for shopping, purchase something online or go to some restaurant, we first check the reviews to make the right decisions. we addressed the task of automatic identification of fake news. We introduced two new fake news datasets, one obtained through crowdsourcing and another one obtained from the web covering celebrities. We can even change to some other better classifier to classify the data other than naïve bayes and logistic regression.

REFERENCES:

1.Majed Alrubaian, Muhammad Al-Qurishi, A Credibility Analysis System for Assessing Information on Twitter, IEEE Transactions on Dependable and Secure Computing, 1-14. DOI : http://dx.doi.org/ 10.1109/TDSC.2016.2602338.

2.Majed Alrubaian, Muhammad Al-Qurishi, A Credibility Analysis System for Assessing Information on yelp, IEEE Transactions on Dependable and Secure Computing, 1-14. DOI : http://dx.doi.org/ 10.1109/TDSC.2016.2602338

3. Granik, M., & Mesyura, V. 2017. Fake news detection using naive Bayes classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON).

4. N. J. Conroy, V. L. Rubin, and Y. Chen, “Automatic deception detection: Methods for finding fake news,” Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.

DOI : 10.36893.DRSR.2021.V09I01N01.834-840