Fake Data Analysis and Detection Using Ensembled Hybrid Algorithm

Dr.K.Mehatha
Computer Science and Engineering
Hindustan Institute of Technology and Sciences
Chennai, India
mentor@hindustanuniv.ac.in

Palagati Bhanu Prakash Reddy
Computer Science and Engineering
Hindustan Institute of Technology and Sciences
Chennai, India
bhanupalagati@gmail.com

Mandi Pavan Kumar Reddy
Computer Science and Engineering
Hindustan Institute of Technology and Sciences
Chennai, India
Kingpavan.mandi@gmail.com

Ganjikunta Venkata Manaswini Reddy
Computer Science and Engineering
Hindustan Institute of Technology and Sciences
Chennai, India
manaswini@gmail.com

***Abstract*—Fake data detection is the most important problem to be addressed in recent years, there is a lot of research going on in this field. Because of its serious impacts on the readers. researchers, government and private agencies working together to solve the issue. This paper represents a hybrid approach for fake data detection using the multinomial voting algorithm. This algorithm was tested against a fake news dataset which resulted in an accuracy score of 94 percent which is a benchmark in the machine learning field where the other algorithms are at a range of 82 to 88 percent. The list of algorithms that have been used here is as follows Naïve Bayes, Random Forest, Decision Tree, Support Vector Machine, K Nearest Neighbours. Based on them the new hybrid algorithm has been created. All these algorithms use training data as the bag of words model which was created using Count Vectorizer. Experimental data has collected from the Kaggle data world. The above results may vary based on the method of implementation. Anyway, there is still room for the development with the help of NLP.**

I. INTRODUCTION

Data or information is the most is the most valuable asset in this century. So, the most important problem to be solved is to evaluate whether the data is relevant or irrelevant. Fake data has a huge impact on the people and organizations that may even lead to the end of the organization or panic the people.

Machine learning scientists believe that this problem can be solved using the machine learning algorithms and there is lot of on-going research in this field which lead to the new branch called Natural Language Processing.

But this classification is not that simple there are lot of challenges to go through in order to succeed. Let's start with few of them machine learning works with the data if you are having huge and clean data then there was a great chance of creating great classifier. In order to create a real time application, the algorithm should be fed with the most recent data. Data is of different sizes so that should be properly cleaned to get better results. The typical work flow of the classification task consists of collecting data, then cleaning, removing stop words, using term frequency inverse document frequency vectorizer or count vectorizer and then train the algorithm by using the above vectorized data. Else people will jump directly to the Deep learning which is time consuming and resource consuming.

But in this hybrid algorithm five algorithms have been taken and then they are trained separately. Then to classify the new data hybrid algorithm will use all the algorithms which have been trained already and then it will predict based on their predictions.

The goal of this research is to analyse whether the use of different algorithms at the same time by creating a complex algorithm is more accurate. Then various scores like precision, recall, and f1_score has been validated. To prove that this is better than the Deep Learning.

The list of algorithms that has been used here are as follows ***Naïve Bayes, Random Forest, Decision Tree, Support Vector Machine, K Nearest Neighbours*.** Based on them the new hybrid algorithm has been created.

II. FAKE DATA PROPERTIES

In order to declare any data as fake data there should be some checks if the data didn't pass those checks then that data will be labelled as the fake data.

- They have lot of similar words in their repository
- They are having some emotional or offer words
- They are mostly opposing the already known facts
- Then content is not acceptable because of their contradictory behaviour

Based on the above statements there are lot of

relations between different spam and fake messages. If people can classify a huge set of messages as fake or not then by studying all those articles, we can classify the other articles as fake or real by using already classified data as training set for the machine learning model.

In the above scenario a human is being replaced by a machine which is perfectly crafted to find the relation between the items. Along with the above statements it may also find lot of other relations which are not understandable by a human being.

III. TRAINING TESTING DATA

Data has been collected Kaggle data world. It has four columns and they are index, title, text and label of the various news articles of various journals. Among the four text is considered as an independent variable and label is considered as a dependent variable. There is no change in the accuracy scores even with the use of titles because the title words will get repeated in the text column.

The train and test data split were in the ratio of 80 – 20 using random function where the train set is used for the training purpose and test set is used for the testing purpose

IV. ANALYSIS OF ALGORITHMS

A. NAÏVE BAYES FOR FAKE DATA ANALYSIS

Naïve Bayes classifier is a simple probabilistic classifier based on the Bayes theorem with great(naive) independence assumption between the data features, where the class labels are drawn from some finite set. It is not a single algorithm to train such classifiers, but a collection of algorithms based on a common principle: every naive Bayes classifier assumes that the value of a particular feature is independent to the value of any other feature, given the class variable

Naïve Bayes is the most opted statistical technique for the models like email filtering, spam filtering and so on.

Naïve Bayes works on the bag of words features where the data of different articles collected is converted into encoded format by using various vectorization techniques based on the requirement some of them are count vectorizer (CV), term frequency and inverse document frequency vectorizer (TFIDF)

The Bag of words will be passed to the Naïve Bayes model as a training data and based on the data the model will learn

Then when any article is passed to classify vectorizer will create sparse matrix and then model will predict based on the word distribution in the sparse matrix.

$$Pr(F|W) = Pr(W|F)\cdot Pr(F)/(Pr(W|F)\cdot Pr(F)+Pr(W|T)\cdot Pr(T)), \quad (1)$$

where:

Pr(F|W) – conditional probability, fake data when the word present in the article;

Pr(W|F) – conditional probability of finding the word W in fake data articles;

Pr(F) – overall probability that the given data is fake data;

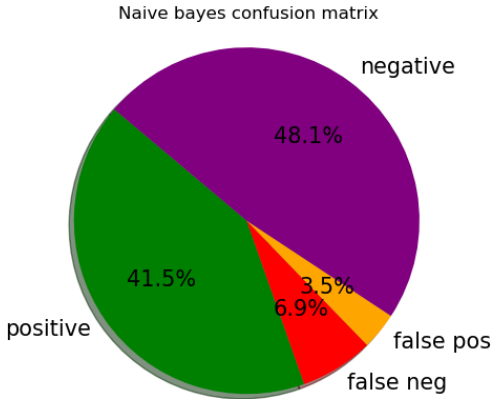
Pr(W|T) – conditional probability of finding the word W in real data articles;

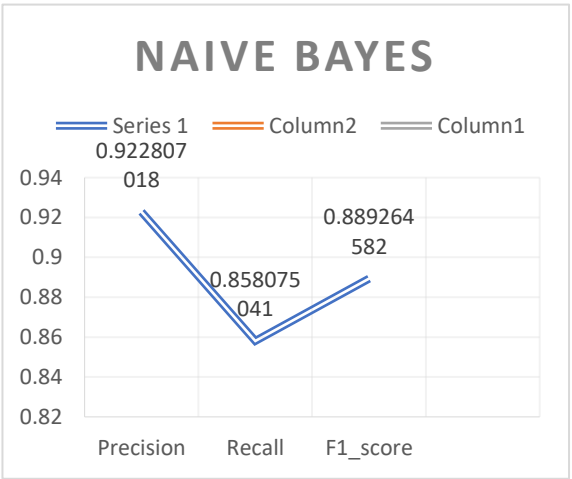
Pr(T) – overall probability that given data is true data. This formula is based on the Bayes' theorem.

Accuracy score achieved by Naïve Bayes algorithm is 0.8966

i. CONFUSION MATRIX VISUALIZATION:

- Confusion matrix is a great way of analysing a machine learning model. This has the data of *True positive, False positive, False negative, True Negative.*
- The confusion matrix of Naïve Bayes is as follows [526, 87, 44, 610]





B. SVM FOR FAKE DATA ANALYSIS

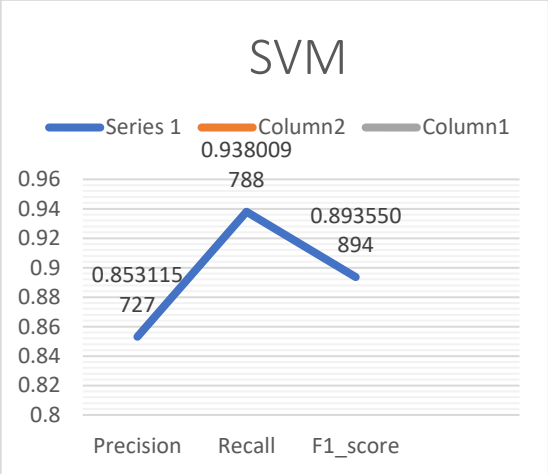
A Support Vector Machine (SVM) is a discriminative classifier where discrimination is achieved by a separating hyperplane. When labelled training data (*supervised learning*) is provided, the algorithm will output an optimal hyperplane which can categorize new examples. The dimensions of the plane are truly based on the number of independent variables in the training data.

SVM is more robust and accurate algorithm compared to the other algorithms over there. This is based on the kernel trick, this is widely used in the distance-based classification tasks.

Support Vector Machine works on the bag of words features where the data of different articles collected is converted into encoded format by using various vectorization techniques based on the requirement some of them are count vectorizer (CV), term frequency and inverse document frequency vectorizer (TFIDF)

Accuracy score achieved by Support Vector Machine algorithm is 0.8918

i. CONFUSION MATRIX VISUALIZATION:
Confusion matrix is a great way of analysing a machine learning model. This has the data of *True positive, False positive, False negative, True Negative.*
The confusion matrix of SVM is as follows [575, 38, 99, 555]



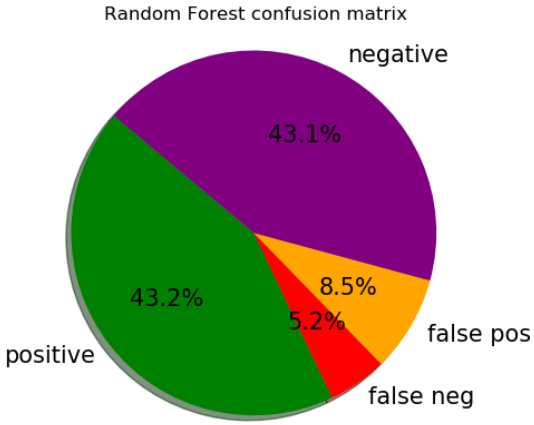
C. RANDOM FOREST FOR FAKE DATA ANALYSIS

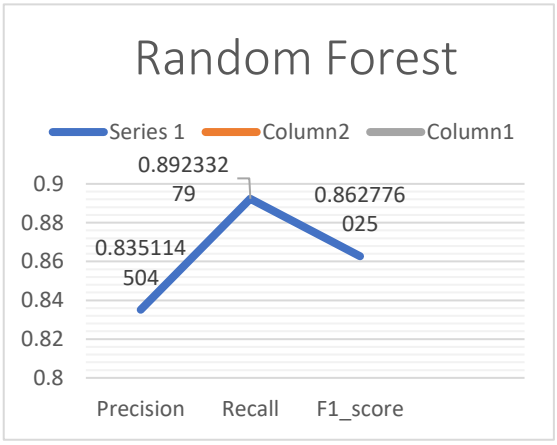
Random forests are ensemble learning method for the classification, regression and some other tasks that are operated by constructing a multitude of the decision trees at the time of training and outputting the class that is one among the training data(classification) or mean predicted value (regression) of the individual trees calculated. Random decision forests are best to solve the overfitting problem of the training data using the decision tree classifier.

Random Forest works on the bag of words features where the data of different articles collected is converted into encoded format by using various vectorization techniques based on the requirement some of them are count vectorizer (CV), term frequency and inverse document frequency vectorizer (TFIDF)

Accuracy score achieved by Random Forest algorithm is 0.8626

i. CONFUSION MATRIX VISUALIZATION:
Confusion matrix is a great way of analysing a machine learning model. This has the data of *True positive, False positive, False negative, True Negative.*
The confusion matrix of Random Forest is as follows [547, 66, 108, 546]





D. KNN FOR FAKE DATA ANALYSIS

In data analysis, the k Nearest neighbours' algorithm (k-NN) is a non-parametric method that can be used for both classification and regression. In either case,

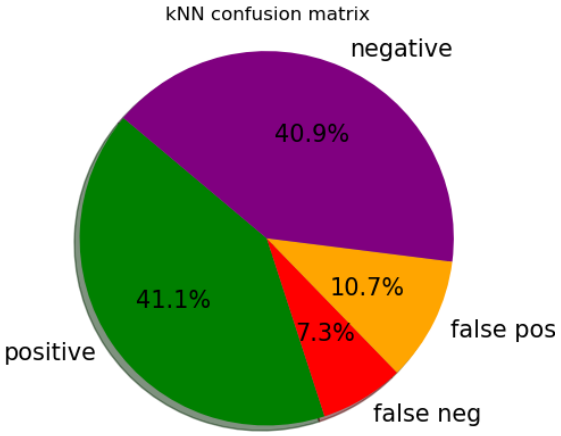
the input consists of the k closest neighbouring training examples in the feature vector. The output will depend on whether k-NN is used for classification or regression task.

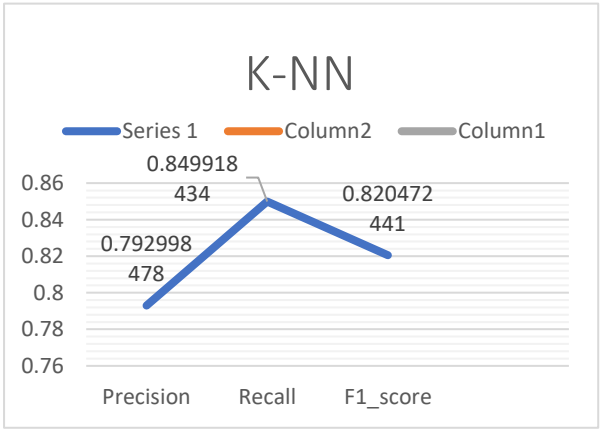
K Nearest Neighbour's, works on the bag of words features where the data of different articles collected is converted into encoded format by using various vectorization techniques based on the requirement some of them are count vectorizer (CV), term frequency and inverse document frequency vectorizer (TFIDF)

Accuracy score achieved by K Nearest Neighbour's algorithm is 0.82

i. CONFUSION MATRIX VISUALIZATION:

- Confusion matrix is a great way of analysing a machine learning model. This has the data of *True positive, False positive, False negative, True Negative.*
- The confusion matrix of KNN is as follows [521, 92, 136, 518]





E. DECISION TREE FOR FAKE DATA ANALYSIS

A decision tree model is a flowchart like structure in which each internal node represents a "test" on an attribute each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root node to leaf node will make the classification rules.

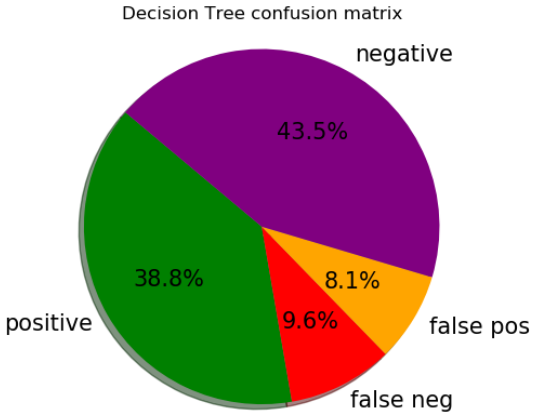
In decision making, a decision tree and a closely related flow diagram are used as visual and analytical decision support tool, where the expected values of competing alternatives are calculated by using the flow.

Decision Tree, works on the bag of words features where the data of different articles collected is converted into encoded format by using various vectorization techniques based on the requirement some of them are count vectorizer (CV), term frequency and inverse document frequency vectorizer (TFIDF)

Accuracy score achieved by Decision Tree algorithm is 0.8279

i. CONFUSION MATRIX VISUALIZATION:

- Confusion matrix is a great way of analysing a machine learning model. This has the data of *True positive, False positive, False negative, True Negative.*
- The confusion matrix of Decision Tree is as follows [495, 118, 97, 557]



V. IMPLEMENTATION OF THE ENSEEMBLED ALGORITHM

The hybrid algorithm proposed is an ensembled algorithm like the random forest which works on the means of the list of algorithms that have been used here is as follows **Naïve Bayes, Random Forest, Decision Tree, Support Vector Machine, K Nearest Neighbours.** Based on them the new hybrid algorithm has been created. All these algorithms use training data as the bag of words model which was created using Count Vectorizer.

By combining all the algorithms, a great change was witnessed in the precision, recall, F1 score, and accuracy values of the test data so this hybrid machine learning algorithm can be used in place of neural network by which time and computational power can be saved.

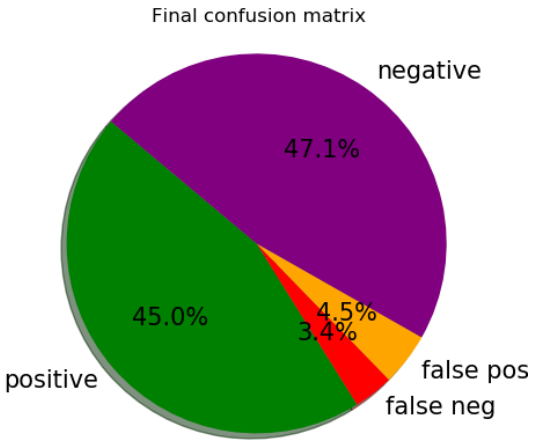
Out (Hybrid) = mode (Out (Naive), Out (SVM), Out (Decision Tree), Out (Random Forest), Out (K-NN))

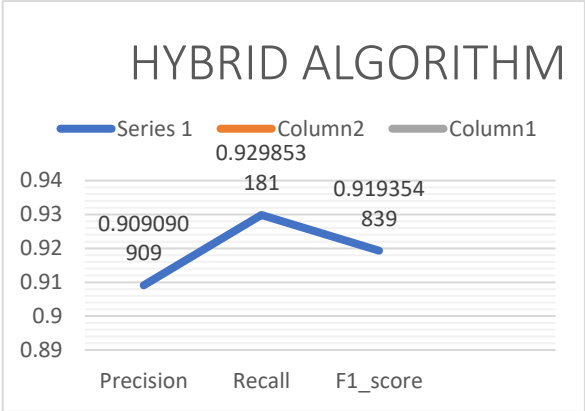
Out (specific algorithm) – output of the specific algorithm for the given test data
Out (Hybrid) – output of the hybrid algorithm
Mode () – implements the mode operation on the outputs

Accuracy score achieved by hybrid algorithm is 0.9258. The overall accuracy is 3 percent more than any other algorithm that has been used for the creation of the multinomial hybrid voting algorithm

i. CONFUSION MATRIX VISUALIZATION:

- Confusion matrix is a great way of analysing a machine learning model. This has the data of *True positive, False positive, False negative, True Negative*
- The confusion matrix of Ensembled Hybrid Algorithm is as follows [570, 43, 57, 597]



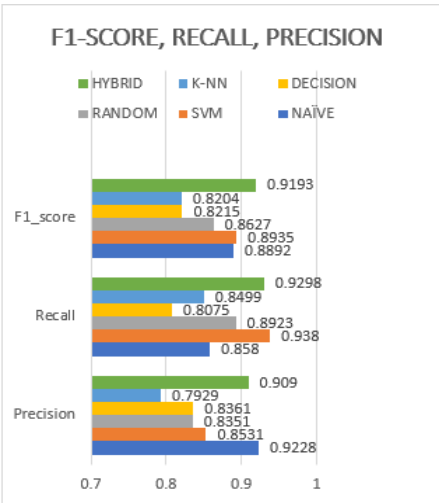


VII. OVERVIEW OF RESULTS

The f1-score, precision, and recall are showcased below. The formulas for calculating them are as below
Precision is defined as the fraction of the relevant instances over the retrieved instances as true.
Precision = True Positive / (True positive + False Positive)
Recall is defined as the fraction of the relevant instances over the total relevant instances.
Recall = True Positive / (True positive + False Negative)
F1-score is based on both the precision and the recall values it is defined as the hormonic mean of the precision and recall
F1-score = (2 * Precision * Recall) / (Precision + Recall)
Both precision and recall are very important for a model because of which F1-score is used as a standard measure for the model comparison

Based on the below bar graph we can clearly pretend that Hybrid algorithm is performing better among all with 91.93% of F1-score

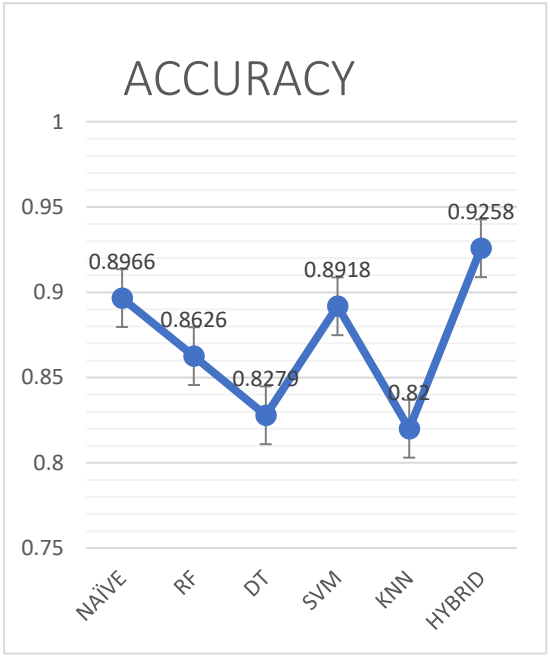
Bar Graph displaying the above three scores



Accuracy scores are yet another important measure for a machine learning model. This score will clearly explain how a model is performing on the test data which is not used for training.

Based on the accuracy graph we can clearly pretend that Hybrid voting algorithm is performing best among all with an accuracy of 92.58

Accuracy scores of different machine learning models



VIII. CONCLUSION

The research showed that ensembled algorithm can perform better when compared with single algorithm in an aspect. By observing confusion matrix, a conclusion can be drawn that various algorithms are learning in various ways some are making less false negatives and some are making less false positives. If various algorithms are trained and a test data is passed by all the algorithms and mode of their results will be announced as the final result gave huge change in the accuracy scores. There is an increase 2.5 percent at least and 10 percent at most. So, the research concluded that ensembled algorithm is far better than a single algorithm.

IX. FUTURE WORK

As a future project this ensembled algorithm will get compared with the deep neural networks and test results will be drawn. If this performs better then lot of time can be saved in training the deep neural networks. This can even change usage of machine learning over the datasets.

X. REFERENCES

1. Fake News Detection Using Naive Bayes Classifier, Mykhailo Granik, Volodymyr Mesyura, Computer Science Department, Vinnytsia National Technical University Vinnytsia, Ukraine.

2. Fake News Detection, Akshay Jain, Amey Kasbe, Department of Electronics and Communication Engineering, Maulana Azad National Institute of Technology, Bhopal, India.

3. Hybrid Machine-Crowd Approach for Fake News Detection, Shaban Shabani, Maria Sokhn, 2018 IEEE 4th International Conference on Collaboration and Internet Computing.

4. Understanding User Profiles on Social Media for Fake News Detection, Kai Shu, Suhang Wang, Huan Liu, Arizona State University, Tempe, AZ 85281

5. A Feature Based Approach for Sentiment Analysis using SVM and Coreference Resolution, Hari Krishna M, Rahamathulla K, Ali Akbar, Department of Computer Science and Engineering, Govt. Engineering College Thrissur, India

6. A Naïve Bayes Approach for working on Gurmukhi Word Sense Disambiguation, Himdweep Walia1, Ajay Rana2, Vineet Kansal3, (1)GNIOT, Gr. Noida, Uttar Pradesh, India, (2)Amity University Uttar Pradesh, Noida, India. (3)CSED, IET, Lucknow, Uttar Pradesh, India.

7. A novel Naive Bayes model: Packaged Hidden Naive Bayes, Yaguang Ji, Songnian Yu, Yafeng Zhang, School of Computer Engineering & Science, Shanghai University Yanchang Rd. 149, 200072 Shanghai, China

8. Accuracy ofClassifier Combining Based on Majority Voting, Peng Hong, Lin Chengde, Luo Linkai, Zhou Qifeng, Department of Automation, Xiamen University Xiamen, P.R.China.

9. An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data Set, Andrew Christian Flores, Rogelyn I. Icoy, Christine F Pena, Ken D. Gorro, Department of Information and Computer Sciences, University of San Carlos Cebu, City, Philippines.

10. Analysis of ordering based ensemble pruning techniques for Voting based Extreme Learning Machine, Sukirty Jain, Sanyam Shukla, Bhagat singh Raghuwanshi, Computer Science and Engineering, MANIT, Bhopal, India.

11. Detecting Depression Using K-Nearest Neighbour's (KNN) Classification Technique.

12. Evaluating Machine Learning Algorithms for Fake News Detection, Shlok Gilda, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India

13. Fake News Detection Enhancement with Data Imputation, Chandra Mouli Madhav Kotteti, Na Li, Lijun Qian, CREDIT Center, Prairie View A&M University, Texas A&M University System, Prairie View, TX 77446, USA.

14. Identifying tweets with Fake News, Saranya Krishnan, Min Chen, Division of Computing and software systems, School of STEM, University of Washington Bothell, USA.

15. Sentiment Analysis Using Multinomial Logistic Regression , Ramadhan WP1, Astri Novianty S.T.,M.T2, Casi Setianingsih S.T.,M.T3, Department of Computer Engineering, Telkom University.

16. Sentiment Analysis using Random Forest Ensemble for Mobile Product Reviews in Kannada, Yashaswini Hegde, S.K Padma, SJCE Mysore.