Social Media Hate Recognition System

By :

Tanushree Gorai(20BCCE1296) . Tanvi Bahedia(20BCE1667)

***Abstract*— This initiative seeks to identify hate speech in Twitter. For ease of use, we define a tweet as having hate speech inside it if it has a racist or sexist emotion. To categorise racist or sexist tweets from other tweets is the challenge at hand.**

Our goal is to predict the labels on the test dataset given a training sample of tweets and labels, where label '1' denotes the tweet is racist or sexist and label '0' denotes the tweet is neither racist or sexist.

We'll use Text Normalization, Vectorization, SMOTE for data pre-processing and Logistic Regression, Naive Bayes Classifier, Random Forest Classifier, Extreme Gradient Boosting Classifier to check accuracy of output.

***Keywords*—** Twitter, SMOTE, Text Normalization, Sentiment analysis (SA), Vectorization, Prediction, Machine learning.

I. INTRODUCTION

Sentiment analysis is the systematic identification, extraction, quantification, and study of affective states and subjective data. It makes use of natural language processing, text analysis, computational linguistics, and biometrics. Sentiment analysis is frequently used in marketing, customer service, and clinical medical applications. It is used to voice of the customer materials including reviews and survey replies, internet and social media, and healthcare materials.

Section 2 describes related work based on a literature review. Section 3 discusses performance analysis and visualization, and Section 4 concludes the paper on the importance of analysis.

A. Problem Statement

Inspite of the fact that there is software available to extract information about a person's opinions of a certain good or service, companies and other data workers still have problems with the data extraction.

- Web-Based Application Sentiment Analysis Concentrate on One Tweet Only - People are using social media sites like Twitter, which produce large volumes of opinion writings in the form of tweets and are available for sentiment analysis, while the World Wide Web expands quickly . From a human perspective, this equates to a vast volume of information, making it challenging to quickly extract sentences, read them, analyse tweet by tweet, summarise them, and organise them into an understandable style .

- Sentiment Analysis Challenges with Inappropriate English - The term "informal language" describes the usage of slang and colloquialisms in conversation, including expressions like "Could not" and "Couldn't." The analysis and decision-making process may be hampered by the inability of some algorithms to recognise sentiment from the usage of informal language.

II. DOMAIN - DATASETS

Training Dataset : train.csv - For training the models, we have a labelled dataset of 31,962 tweets. The dataset is provided in the form of a csv file with each line storing a tweet id, its label and the tweet.

Test Dataset : The test data file contains only tweet ids and the tweet text with each tweet in a new line.To test the accuracy of our model and check its validity we have taken a count of 17,198 tweets just to achieve a higher rate of accuracy of our model.

III. LITERATURE SURVEY

A. Sentiment Analysis of Twitter Data: A Survey of Techniques - Vishal A. Kharde and S.S. Sonawane.

In this research, they present a survey and comparative analysis of the existing methods for opinion mining, such as lexicon-based and machine learning approaches, cross-domain and cross-lingual techniques, and some assessment metrics. The findings of the research indicate that lexicon-based approaches are sometimes highly effective and require little effort in human-labeled documents, while machine learning methods, such as SVM and naive Bayes, have the highest accuracy and can be regarded as the baseline learning methods. We also looked into how different features affected classifiers. We can draw the conclusion that more accurate results can be attained with cleaner data. When compared to other models, the bigram model's use gives superior sentiment accuracy. In order to increase the precision of sentiment classification and their ability to adapt to a wide range of domains and languages, they concentrated on the study of combining machine learning methodology with opinion lexicon methodology.

B. Twitter Sentiment Analysis - *Aliza Sarlan, Chayanit Nadam, Shuib Basri*

Twitter sentiment analysis is a tool created to examine client perceptions of things that are essential for business success. The application will combine natural language processing methods with a machine-based learning approach, which is more accurate for sentiment analysis. As a result, the program's sentiment will be divided into good and negative categories, and this will be depicted in a pie chart and HTML page. Despite the fact that it was intended for development as a web application, Django can only be used with Linux servers or LAMP. It cannot be realised as a result. Thus, it is advised that this element be further improved in subsequent research.

C. Twitter Sentiment Prediction - *Faizan, Sharda University*

They have demonstrated a system in this study for the analysis of textual twitter data, specifically for the newly developed discipline of sentiment analysis. They created a model for the study of feelings using the KNN method and features from unigram, bigram, and ngram. This model was trained and tested using the #USairline data set, and it achieved an accuracy of 65.33%. They want to carry out similar study in the near future to improve the model's accuracy without extracting any information, utilising other deep learning approaches like neural networks.

D. Sentiment analysis of twitter data - *Hamid Bagheri and Md Johirul Islam, Computer Science Department Iowa State University*

They examined the value of social news analysis and its applicability in several fields in this technical article. They concentrated on Twitter and utilised a Python software to apply sentiment analysis. They displayed the results alongside various current events. They came to the conclusion that the neutral feelings are disproportionately high, demonstrating the necessity to enhance Twitter sentiment analysis.

E. Sentimental Analysis of Twitter Data with respect to General Elections in India

- *Ankita Sharmaa and Udayan Ghoseb ,USICT, Guru Gobind Singh Indraprastha University, New Delhi, India*

The two areas of text mining and sentiment analysis, which are often studied separately, are combined in this study. This paper's methodology was sensible and organised. Prior to conducting sentiment analysis on the tweets, it includes tweet text analysis. The primary and sizeable data set used in the paper is. Based on a thorough literature review, the tools used in this paper were chosen. It is well known that there is no set format for a tweet; the only restriction is on how long the content can be. Since the tweets were gathered from an internet source, no sentiment descriptors were included. The document became more complicated as a result of the sentiment labels' absence and the tweets' clamorousness.

F. Text Normalization

Any voice and language processing application must have text normalisation. Throughout the previous few decades, a lot of works have been published to further text normalisation. For the same purposes, numerous new models and approaches are created for processing various languages. However every system still has significant shortcomings that prevent it from appropriately classifying each NSW. Concatenative synthesis is the most often utilised synthesis for speech synthesis because it produces speech that is natural and understandable. Even after so many attempts, there is still much room for advancement, particularly in the text normalisation stage of speech and language processing.

G. Vectorisation

Word embeddings and word vectorization are texts or strings that are converted to real numbers for Natural Language Processing. Once words have been transformed into vectors, the approach of cosine similarity is utilised to satisfy the majority of use cases for NLP, document clustering, text classification, and word prediction based on sentence context. Cosine Similarity — "Smaller the angle, higher the similarity Word vectors can be converted with the use of well-known designs like Word2Vec, Fasttext, and Glove, which also use cosine similarity for word similarity characteristics. Regarding the enormous word dataset, NNLM and RNNLM perform better. But the cost of sophisticated computations is significant. Word2Vec uses CBOW and Skip-gram architecture to maximise accuracy and reduce computation complexity in order to get around the problem of computation complexity.

H. A Text Normalization Method for Speech Synthesis Based on Local Attention Mechanism

This paper suggests a model LATN for the TTS text normalisation job utilising the encoder-decoder architecture and GRU recurrent network and introduces a local attention technique to retrieve the most crucial context, which enhances text normalisation accuracy. This model motivates us to build a customised model tailored to the particulars of the work in order to improve outcomes and lower computational costs.

I. SMOTE: Synthetic Minority Over-sampling Technique

They have developed an innovative method for handling severely unbalanced datasets in this work called A-SMOTE, which is an enhancement over SMOTE. A-performance SMOTE's was assessed using 44 datasets with high unbalanced classification ratios. Using an ML algorithm, the proposed approach was contrasted with several hybrid oversampling and undersampling strategies (e.g., C4.5, Naive-Bayes). The A-SMOTE technique for preparing imbalanced datasets obtained a greater accuracy and F-measure in our experimental results (F-value). Since it produces high-quality data, the proposed A-SMOTE can be a beneficial tool for researchers and practitioners. Future research will concentrate on the application of A-SMOTE and rough set theory to the categorization of imbalanced datasets.

J. Synthetic Minority Over-sampling Technique - *N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer*

It is described how to create classifiers from datasets with imbalances. If the categorization categories are not roughly equally represented, a dataset is unbalanced. Real-world data sets frequently contain a large percentage of "regular" cases and a very small number of "abnormal" or "interesting" examples. The cost of misclassifying an abnormal (interesting) example as a normal example is also true, and it is frequently much larger than the cost of the opposite error. A useful way to improve a classifier's sensitivity to the minority class has been suggested: under-sampling the majority (normal) class.

IV. METHODOLOGY USED:

A. Data collection :

Data collection is the very first stage in any form of analysis, whether it be technical or not. We need a lot of data to perform analysis on a given problem, and then we use a variety of techniques and algorithms to draw our desired conclusions from the data. It is advisable to collect a lot of data because the more data that is analyzed, the more accurate the results will be and the more confident you can be in the decisions you make based on those results.

B. Feature Engineering : Feature engineering is the pre-processing step of machine learning, which is used to transform raw data into features that can be used for creating a predictive model using Machine learning or statistical Modelling. Feature engineering in machine learning aims to improve the performance of models. In this topic, we will understand the details about feature engineering in Machine learning. We'll try to extract a few characteristics of the tweets to help us separate negative tweets from positive ones.

C. Data pre - preprocessing:

Data processing comes next after data collection. Raw data is information that has been taken straight from a dataset or other data source. The dataset includes a number of fields, including Age, Gender, and others, some of which include null values, which leads to problems in the final product, which is the graphical visualisation of the data. These null values must be removed or replaced with other valid values in order to correct the issue and produce accurate results. To finish this assignment, we employed a process called Deterministic Imputation. In a condition known as "deterministic imputation," the null values (NA or NaN) are computed using the data from the other values in the same column. There are many models available for this purpose, including the Basic Numeric Imputation Model, in which the null value is changed to the Mean or Median of other values in the same column of the dataset.

- **Text Normalization :** Text is normalised in an effort to lessen its unpredictability and bring it closer to a predetermined "standard." By doing so, we can decrease the variety of data that the computer must process, which increases productivity. Lemmatization and stemming are examples of normalisation procedures that aim to reduce a word's inflectional forms and occasionally derivationally related forms to a basic form that is shared by all words.

- **Vectorization :**The term "vectorization" refers to a traditional technique for transforming input data from its original text-based format into real-number vectors, which is the format that is supported by ML models. This method has been around since the invention of computers, it has proven extremely successful across numerous disciplines, and it is now utilised in NLP. Vectorization is a phase in the feature extraction process in machine learning. By translating text to numerical vectors, the goal is to extract some distinguishing features from the text for the model to train on.

- **SMOTE :** One of the most popular oversampling techniques to address the imbalance issue is SMOTE (synthetic minority oversampling technique). By increasing minority class samples at random and duplicating them, it seeks to balance the distribution of classes. SMOTE creates new minority instances by combining minority instances that already exist. For the minority class, it creates virtual training records using linear interpolation. For each example in the minority class, one or more of the k-nearest neighbours are randomly chosen to serve as these synthetic training records. Following the oversampling procedure, the data is rebuilt and can be subjected to several categorization models.

D. ML Modeling

A programme that uses a dataset that has never been seen before to detect patterns or make choices is known as a machine learning model. Machine learning algorithms, for instance, can analyse and accurately identify the intent underlying previously unheard utterances or word combinations in natural language processing.• Histogram• Bar Graph• Box Plot• Scatter Plot

E. Hyperparameter Tuning

A mathematical model with a number of parameters that must be learned from the data is referred to as a machine learning model. We can fit the model parameters by using existing data to train a model.

Hyperparameters, on the other hand, are a different class of parameter that cannot be directly learned through routine training. Usually, they are fixed before to the start of the program itself. These parameters describe crucial model characteristics including complexity and learning rate.

Model hyperparameters include, for instance:

- L1 or L2 regularisation, which serves as the penalty in the Logistic Regression Classifier
- The speed at which a neural network learns.
- The C and sigma support vector machine hyper parameters.
- The letter K as in k-nearest neighbours.

V. ALGORITHM USED

A. Logistic Regression

The logistic regression is the kind of ML modelling where it considers the probability of an event whether it is going happen or not based on the given dataset of independent variables and then concludes its prediction for the same.Since the outcome for the above such scenario the accuracy of the model lies between 0 and 1.

B. Naive Bayes Classifier

It is an algorithm that is based on the famous theorem of probability i.e. Baye's Theorem. It makes the assumption that the features that are used as the input functions or the input dataset are dependent on each other which isn't true in case of the real life world problems which we tend to face in our day to day life.its used to predict the probability of an input which belongs to a particular or a certain class.

C. Random Forest Classifier

Random forest classifier is a kind of ensemble learning method where the main motivation of ensemble learning where the main method used in this is that it combines two or three models together to predict a certain output.Now when we see that there are already three to four models combined hence automatically the accuracy of the model is increased combined to the other models which consider only one algorithm at a particular moment.Similarly in the above method we combine various weak models together to predict the accuracy on a particular dataset where we implement it for the same.

D. Extreme Gradient Boosting Classifier

It is the extension of the gradient boosting algorithm and particularly used for solving complex problems. XGBoost iteratively using various decision trees and by using each tree it tends to correct its mistake in the present tree. At each step the model calculates the gradient and the hessian of the loss function with respect to the predicted values and then fits a decision trees to the negative gradients of the loss function.

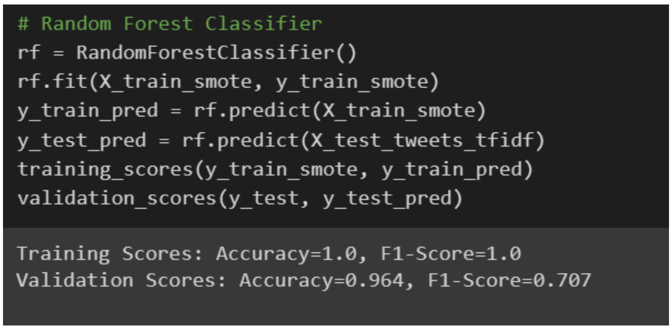
VI. INFERENCE AND OUTPUT

The inference of the paper could be divided into four main categories which is overall analysis of the project, trend analysis of the data, and the comparative analysis.

A. Overall Analysis

The overall analysis of the paper tells us that we were able to collect a dataset of 49000 tweets which consisted of both

positive as well as the negative tweets and our model was trained in such a way that it could predict out the same with an accuracy of 0.964 and with an F1 score of 0.707 which is a better and a higher score than evaluated using some other methods.

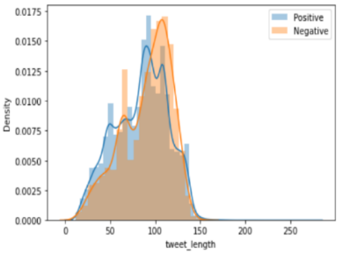


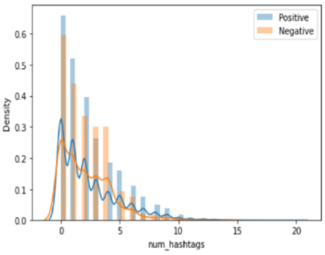
```
# Random Forest Classifier
rf = RandomForestClassifier()
rf.fit(X_train_smote, y_train_smote)
y_train_pred = rf.predict(X_train_smote)
y_test_pred = rf.predict(X_test_tweets_tfidf)
training_scores(y_train_smote, y_train_pred)
validation_scores(y_test, y_test_pred)
```

```
Training Scores: Accuracy=1.0, F1-Score=1.0
Validation Scores: Accuracy=0.964, F1-Score=0.707
```

B. Trend Analysis

Using the techniques of the data visualization we were able to go for a trend analysis in terms of the tweet length compared with the its density with both kinds of tweets which is the positive as well as negative tweets and it gives us a wide picture regarding the same. The next visualization which we were able to compare was the number of hashtags being used similarly for both kinds of tweets And in the similar sense we used all the possibilities of the comparison with its uses as in the exclamation mark, question marks ,number of tags used or words used in a particular tweet.



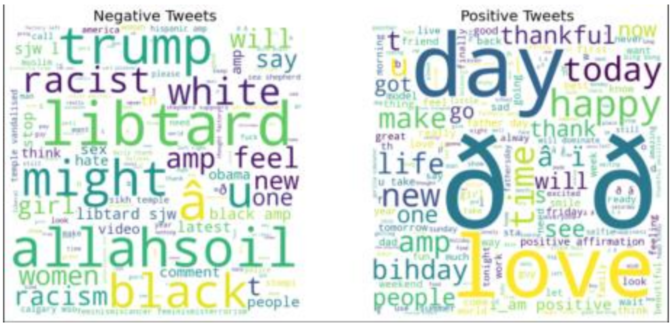


C. Comparative Analysis

Twitter sentiment analysis is the process of analyzing the sentiment of tweets and categorizing them into positive, negative, or neutral sentiments. There are various approaches to perform sentiment analysis, including rule-based, machine learning-based, and hybrid methods. Here's a comparative analysis of these three approaches:

Rule-based approach:

The rule-based approach involves creating a set of rules to classify tweets based on their sentiment. The rules are usually created by analyzing the language used in the tweets, and then assigning a sentiment label based on predefined rules. This approach is relatively simple and doesn't require a large dataset for training. However, it may not be as accurate as other methods because it relies on a limited set of predefined rules.

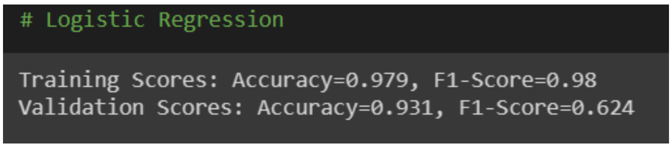


Machine learning-based approach:

The machine learning-based approach involves training a machine learning model using a large dataset of annotated tweets. The model learns to classify tweets based on their sentiment by analyzing the features of the tweets, such as the words used, the frequency of words, and the context in which they are used. This approach is more accurate than the rule-based approach, but it requires a large dataset for training, and the model may not perform well on new and unseen data.

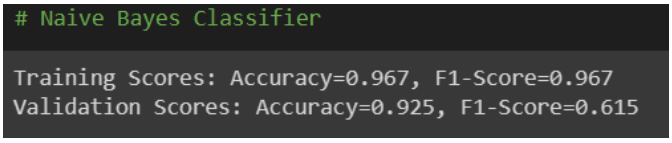
Hybrid approach:
The hybrid approach combines both rule-based and machine learning-based approaches to improve the accuracy of sentiment analysis. The rule-based approach is used to create a set of initial rules, which are then refined and optimized using machine learning techniques. This approach can achieve higher accuracy than either approach alone and can adapt to new and unseen data.

Here are the output for the various techniques we have used for our project.



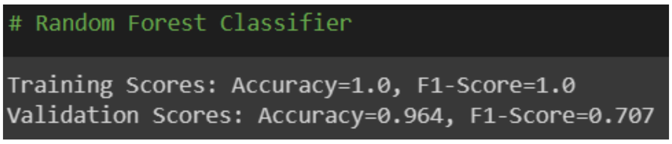
```
# Logistic Regression

Training Scores: Accuracy=0.979, F1-Score=0.98
Validation Scores: Accuracy=0.931, F1-Score=0.624
```



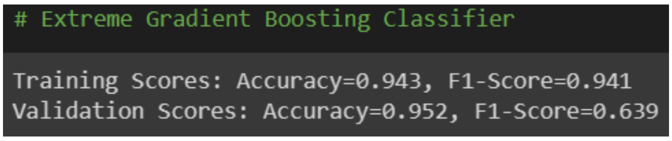
```
# Naive Bayes Classifier

Training Scores: Accuracy=0.967, F1-Score=0.967
Validation Scores: Accuracy=0.925, F1-Score=0.615
```



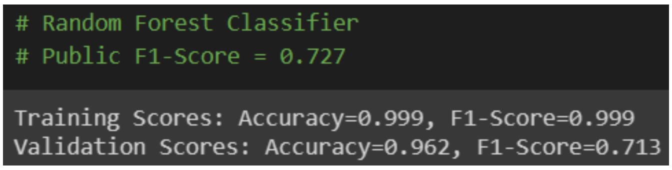
```
# Random Forest Classifier

Training Scores: Accuracy=1.0, F1-Score=1.0
Validation Scores: Accuracy=0.964, F1-Score=0.707
```



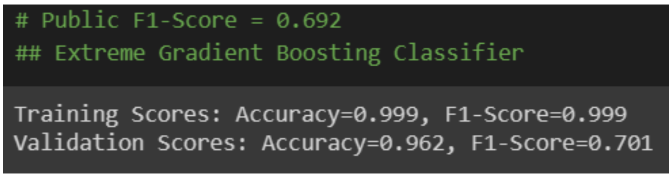
```
# Extreme Gradient Boosting Classifier

Training Scores: Accuracy=0.943, F1-Score=0.941
Validation Scores: Accuracy=0.952, F1-Score=0.639
```



```
# Random Forest Classifier
# Public F1-Score = 0.727

Training Scores: Accuracy=0.999, F1-Score=0.999
Validation Scores: Accuracy=0.962, F1-Score=0.713
```



```
# Public F1-Score = 0.692
## Extreme Gradient Boosting Classifier

Training Scores: Accuracy=0.999, F1-Score=0.999
Validation Scores: Accuracy=0.962, F1-Score=0.701
```

VII. REFERENCES

1. Hamid Bagheri and Md Johirul Islam - Computer Science Department Iowa State University - "Sentiment analysis of twitter data "

2. Yili Wang , Jiaxuan Guo , Chengsheng Yuan and Baozhu Li - "Sentiment Analysis of Twitter Data"

3. Vishal A. Kharde and S.S. Sonawane - Sentiment Analysis of Twitter Data: A Survey of Techniques - International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016

4. Aliza Sarlan, Chayanit Nadam, Shuib Basri - Twitter Sentiment Analysis - 2014 International Conference on Information Technology and Multimedia (ICIMU), November 18 – 20, 2014, Putrajaya, Malaysia

5. Abdullah Alsaeedi and Mohammad Zubair Khan - "A Study on Sentiment Analysis Techniques of Twitter Data" -(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2, 2019

6. Faizan - Sharda University - "Twitter Sentiment Analysis" - Volume 4, Issue 2, February – 2019 International Journal of Innovative Science and Research Technology

7. Ankita Sharma and Udayan Ghose - "Sentimental Analysis of Twitter Data with respect to General Elections in India" - International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020

8. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau - "Sentiment Analysis of Twitter Data" - Department of Computer Science Columbia University New York, NY 10027 USA

9. Pooja Manisha and Rahate, Manoj Chandak - "An Experimental Technique on Text Normalization and its Role in Speech Synthesis" - International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8S3, June 2019

10. Prabhu - "Understanding NLP Word Embeddings — Text Vectorization" - Published in Towards Data Science, Nov 11, 2019

11. Lan Huang, Shunan Zhuang and Kangping Wang - "A Text Normalization Method for Speech Synthesis Based on Local Attention Mechanism" - February 2020 IEEE Access PP(99):1-1, DOI:10.1109/ ACCESS.2020.2974674, License : CC BY 4.0

12. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer - "SMOTE: Synthetic Minority Over-sampling Technique" - Submitted on 9 Jun 2011- Journal Of Artificial Intelligence Research, Volume 16, pages 321-357, 2002

13. Ahmed Saad Hussein, Tianrui Li, Wondaferaw Yohannese Chubato, and Kamal Bashir - "A-SMOTE: A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE" - November 2019, International Journal of Computational Intelligence Systems

14. Jeff Heaton - "An Empirical Analysis of Feature Engineering for Predictive Modeling" - arXiv:1701.07852v2 [cs.LG] 1 Nov 2020

15. Bhartendoo Vimal - "Application of Logistic Regression in Natural Language Processing" - June 2020, International Journal of Engineering and Technical Research V9(06)

16. Martiti and Christina Juliane -"Implementation of Naive Bayes Algorithm on Sentiment Analysis Application" - Advances in Engineering Research, volume 207, Proceedings of the 2nd International Seminar of Science and Applied Technology (ISSAT 2021)

17. Bahrawi - "SENTIMENT ANALYSIS USING RANDOM FOREST ALGORITHM-ONLINE SOCIAL MEDIA BASED" - a journal of information technology and its utilization, volume 2, issue 2, December - 2019

18. Rudra Sarker Utsha1 , Mumenunnessa Keya , Md. Arid Hasan and Md. Sanzidul Islam - Daffodil International University, Dhaka, Bangladesh - "Qword at CheckThat! 2021: An Extreme Gradient Boosting Approach for Multiclass Fake News Detection"

19. Adele Cutler, D. Richard Cutler and John R. Stevens - "Random Forests" - January 2011 Machine Learning 45(1):157-176 , DOI:10.1007/978-1-4419-9326-7_5

20. Philipp Probst, Anne-Laure Boulesteix and Bernd Bischl - "Tunability: Importance of Hyperparameters of Machine Learning Algorithms" - Journal of Machine Learning Research 20 (2019) 1-32 Submitted 7/18; Revised 2/19; Published 3/19