Turkish Journal of Computer and Mathematics Education *Vol.12 No.3(2021), 4918-4926*

Research Article

Detecting Misleading Information on COVID-19 : A Machine Learning Perspective

M.Arunkrishna[a,*] , B.Mukunthan[b]

[a]Research scholar, Department of Computer Science, Jairams Arts and Science College(Affiliated to Bharathidasan University, Tiruchirappalli),Karur - 639003, Tamilnadu, India.
[b]Research Supervisor, Department of Computer Science, Jairams Arts and Science College (Affiliated to Bharathidhasan University), Karur - 639003, Tamilnadu, India.
([*]Corresponding author's e-mail: arunkrishna.murugan@gmail.com)

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: Online Social networks become a popular way for sharing information among people. With increasing technology like Wi-Fi, Wi-Max ,3G/4G along with handheld devices like smartphones and tablets, popular applications such as Instagram, Facebook, Twitter and YouTube, becomes a dominant platform for news and entertainment. The extensive use of these social networks has an incredible influence on sharing news among people It holds both positive and negative effects of its own. Because of it's high popularity,Online Social Networks(OSNs),has become the target for spammers. Also, false news for different political and commercial purpose has been evolving in the large count and spread worldwide. After the spread of COVID-19, there had been a lot of confusion and pitfalls on the topic of who to believe and who should be rejected. With the advent of time, several companies like Facebook, and Twitter joined hands to identify the news and regard it authentic or not. This effort was very hard for people, as the news are spreading at a rapid pace, no matter how many people are upon the task, the rate of expansion of news is always faster than the rate of evaluation of whether the news is authentic or not. Additionally, it can be observed that the news cannot be regarded as fake or true before careful evaluation. This evaluation is based on the results. So it is important to create a method for identifying fake news and distinguishing it from individuals. Thus, the paper evaluates several models in order to find the best fit with the highest level of accuracy..

Keywords: Spam Detection, Machine Learning, NLP, COVID- 19 spam,False news,misleading information

1. Introduction

Novel corona virus also known as COVID-19 is a type of virus (SARS-CoV-2) which started to spread from Wuhan, China during December 2019. World Health Organization (WHO) announced that, COVID-19 virus infection as a public health emergency. While this issue, misinformation related to novel corona virus has also flooded in the internet and caused severe social disruptions as well. To help people combat this misinformation, researchers are working on some computational methods that automatically detect, debunk and mitigate diverse types of false information. News of all sorts come into the system, and this news never makes it to the control room for evaluation. In the context, there were several factors that are considered, and if there were a filter which could take in all sorts of news, and then evaluate the parameters, in order to determine the prospects, this filter would highlight which news is fake and which one is original. This further state that the prospects make people believe. As the model would continue to be the predictor of the authenticity of the watch.

Additionally, for this purpose, various models were proposed, but seeing the dataset considering, whether the website is authentic, or the text and the title. These are the measures where the news is to be recorded. Keeping in view these factors, the best and optimal solution for the dataset and its accurate prediction is based on Natural Language Processing in terms of the algorithm that would identify the level of affiliation of words with the importance of the whole mechanism. This involves the preparation of algorithms and responses that not only highlight which words were used but also indicate their connection and affiliation with the resultant classifier, based on the values that are presented dataset.

In terms of the dataset, this is an evaluation in a data frame, this not only makes the potential of the model higher but also indicates that the dataset has a total of four variables, the first variable is the source of information, the second variable is the title of the argument, after the title, comes the article text. It illustrates the level of connectivity; the last variable is the classifier identifying whether each piece of news. This variable indicates that the potential prediction of the variable and takes it to the next level, where this potential would enable the model to predict, whether the news is accurate or not, with high accuracy, and reliability, without looking for some human interference in them.

2.Literature Review

In terms of the past studies, on NLP and news identification based on the content of the writing, there are several factors that should be loaded, and this would also enable the overall model to be a basic predictor, based on only its content. This makes the research unique and outstanding as it would not be using any other references to mark the correct and wrong networks, and then lead the resources to a prediction. There are several factors that should be included in an NLP research [1], as studies the potential of the content to be a dominant predictor, where it could make amendments to the current model, and further, be seen that using NLP technique can result in the creation of an algorithm, that would increase the number of features in a model [2]. This would lighten up the regional contexts that change with the changing level of production. Moreover, this tends to lead to the supported conclusion that both sentiments and actuation is a measure of the content [3]. It was knowing this and the potential variable that not only makes the top of the news that makes up for the classification engine that limits the flow of these structures [4]. Therefore, the classification should be based on the decision tree, and creating the overall analysis that makes up for the model.

3.Methodology

The methodology or the base model that is used in this study is the random forest.This is one of the best models used for classification for NLP purposes [5,6]. As it is fast and convenient that not only limits the timeframe but also understands the accuracy to be one of its key factors and the natural state to be predicted in layers, but the prediction is highly accurate [7,8]. Moreover, there are two key metrics that were used in this model, these were the number of estimators, and the depth of the model [9]. This not only concluded the variables but also illustrated that the model was to pass a total of 5 tests. The method for this optimization was based on the grid search. This grid search ran all the variations of the model in terms of its hyperparameters and measured which variation as best in terms of the model [10].

Additionally, all the models were compared graphically, and the best model in terms of each metric was not only identified, but they were explained. This led to the results of higher variation and lower synchronization rates enabling to work real-time and identify which articles or news are real and which ones are not. This would be done by an ensemble model, random forest, in terms of a grid search with different hyper-parameters. The process that was followed in the preparation of the dataset for machine learning, and its further analysis is discussed in detail as follows.

3.1 Selection of Dataset

This was the first and foremost step, as the objective was to use machine learning capabilities to mine the data of COVID-19, in terms of natural language processing, and this accounts not only for the valuation of which news is a fake and which one is real. This also accounts for the recognition of a relationship between the content of the essay and its validity, with respect to Corona. Therefore, a sample size labelled as fake and real was converged, and tallied, and the news was tallied, if these were actual, and not misinformation, then it was marked, so, and if the objective and aim was just marketing and not facts, then it was regarded as misinformation.

Dataset Identification

The dataset was identified and was pre-recorded, and this made the potential of the dataset illustrative. Moreover, this made the potential of the dataset exquisite, as it had a timestamp, a title, that illustrated what the article is all about, and then there was the content and the source. These were the variables in the study, and the last variable was a label, that identified whether the results are real or fake. Moreover, it can be understood that the potential factors that the dataset implores are based on the factor that resides in the processing of natural language.

Expected Features

Though in the original dataset, without pre-processing there were four, and if the label was included in it, there were a total of five variables. Moreover, it can be illustrated that the factors that accounted for most are the source, the title and the text also termed as the content of the complete essay. Moreover, it is up to the algorithm to identify which variables are original and which ones are not. There is an extensive pre-processing and loading face in terms of the data preparation for the given model.

3.2 Loading and Exploration of Dataset

The data was loaded into python via a csv, and it took the data frame to evaluate the need and necessity to be loaded in that format. The loading and exploration of the dataset were done in this phase. Moreover, the overall analysis was performed, and this analysis indicated that the results of starting the data analysis were based on the division of data. Additionally, loading and exploration were divided accordingly.

In terms of loading the dataset, the csv was loaded into Pandas data frame by python, this had a total of 5 variables, out of which four of them were features, and one was a label, for the record, the dataset was not divided into classes in this model. Moreover, the results were evaluated, and the analysis was performed on the basis of the model. This indicates that the potential of the variable tends to be in term of the complete dataset.

In the exploration part of the dataset, there were five variables that were studied. These variables indicated different trends. At the same time, the content of each paper was unique, though some were missing, still applicable with the combination of a title and the source. Moreover, there were factors that accounted for in this dataset, with the source being one of the dominant variables that are both nominal and has some repetitions. Moreover, there was the label that was the binomial variable, indicated either the dataset row was a fake or it was real. This was the predictor, and the model had to predict this accurately.

3.3 Organization of dataset

This phase was to organize the dataset, and it incorporated the cleaning of the dataset and making it viable for the model, as, on the other hand, it would be a sole prediction from the source, and it is common sense, that the source does not always prevent misinformation, though there are chances that the factors involved would be most important, the objective of this is to predict via Natural Language Processing (NLP). The organization of dataset was based on the following metrics, and it incorporated the measures necessary to start the feature making of the model,

Removing Punctuation

While punctuation are an important part of a sentence and are vital for humans to understand its meaning, it is a secondary element of a sentence according to the machine learning models. Moreover, there are factors that account for this change, and the factors also make it possible to evaluate a sentence without its punctuation. So, in this part of the organization, the factors that account for most are the words, not the punctuation, so in this step, all the punctuation were removed from the content and the title, and this enabled a direct, approach for the next steps. Though for humans a sentence without punctuation would be a body without a soul, for a computer algorithm, punctuation simply indicates the expression that confuses both the model and the computer and tends to give real results.

Tokenization

Tokenization is the act of separating the content into either sentences or words. This creates lists of words, in this case, that would not only help the separate words but also identifies them in terms of tokens for future analysis, helps the identification of words in the content. For the purpose of this essay, tokens were used for both titles and text. Moreover, there were factors that were incorporated to identify the tokens and their general rule. Tokenization further indicates the number of words in each news factors; this would help indicate how long this text is in terms of words, another reason for tokenization is to remove the stop words from a given text. This becomes easier if the content is tokenized.

Removing of stop words

Stop words are the words inside each content that is used to explain to the user whether there is a need for analysis, or it is an essential stop word. This is termed a stop word, as it does not contribute to the subject at hand, instead, it provides no insight into the problem, but it remains a part of the speech. Removing these stop words enables the computer and the algorithm to remove all useless information and incorporate only that part of the information that presents a real meaning. In this part, as it can be seen, that both the title and the.

3.4 Pre-processing of dataset

After the organization of dataset, the next step that was completed in order to make a natural language processor was to pre-process the data. Since all the data is in the form of lists now and can be evaluated now, the factor that was included in the test was its pre-processing and making it applicable for the model. Moreover, the model indicates that the potential variable is primarily both the title of the article and its content, so both of these should be incorporated, and the result should be taken into consideration. Hence there should be pre-processing in terms of the following steps.

Stemming

This is the first pre-processing stage, and it not only simplifies the potential variation of similarly themed words but also reduces the repetition. Words such as the continuous words with an “ing” at the end, or adverbs with “ly” at their end, there are factors that should be incorporated, and even the plurals with “s” at the end are all stemmed words, and objecting, objected, and even objects should be considered simply object as this would be easier to train. For this model, stemming of both the titles and the texts was done, and this indicated that the

variation was dependent upon the factor of change. This not only reduced the number of features but also made the model more appropriate and directed towards a singular goal.

Lemmatizing

This the second pre-processing of data included the factors that matter most, while there were stemming of the given words, lemmatizing was also done to make things clearer for both the title and the text. Thus, this indicates a more dictionary friendly word and would illustrate if the word or its dictionary equivalent was available in the text. This also proved useful in terms of the next steps, but for the purpose of further study, lemmatized words were taken as an example and introduced into the subdomain of pre-processing.

3.5 Vectorization of Dataset

Once the complete list of stemmed or lemmatized words was acquired, the next phase was the vectorization. This is an important step, and it creates a feature for each word, this indicates that the dataset is getting ready to be put inside a machine learning model. This would not only help the model to precisely considering the weight of each word but would also illustrate the degree to which the potential of those words was to be taken seriously. In other words, the importance of each word in the title and in the text would be the result of this model.

Bag of Words

This is the first approach to the vectorization, and this illustrates the words that are present in the content of each article to process, and after that a new feature is created for each word, making a dummy variable, if the word is present in an essay, then this would put 1 in the row, under the column of the word. If it is not present, then there is a zero it, this is called word vectorization, or the bag of words, it does not tell the number of times the word has occurred but indicates all the processable words that have occurred in the content. For this task, the bag of words or count vectorizer was performed both on the title and the text. This enabled both the words to be combined, and indicate which ones had titles and where was the text. Though the bag of words for title and text were not combined, in order to give a greater overview, but the presence indicated that these words were present in either the title or the text, or even both.

N-Grams

N Grams is the combination of adjacent words; this helps the vectorization of the words. This technique highlights the continuity and consistency of the topic, and also manages to give the rating, as if there were no adjacent words, this would be a unigram. If there were two words that are adjacent, then it would be a bigram, and if there happen to be three words with continuity, then the potential variation and the variability in terms of the N-Grams would be three or trigram, and this process continues. Thus, this creates a new feature that measures the grams of the text, and the longest chain counts, for the most part, not considering the shorter of the two mixes. Moreover, this new variable or feature is a categorical one and indicates whether the content for both the title and the text is created and incorporated for the sake of analysis. This assumes that unigrams have lesser details than bigrams or trigrams, and the higher the chain of words, the greater would be its depth. Additionally, with the increase in grams, there would be more context to the given problem, and therefore the article tends to be more authentic, than unigrams or even bigrams.

TF-IDF

TF-IDF is also termed as the relative frequency, and this indicates for each variable, the frequency of the word in a given content versus the total frequency in the complete dataset. This is calculated for every word and is normally in terms of percentage. Moreover, it can be observed that the factor that should be incorporated in this are the relative importance of each word. If a word occurs more in the content of one article has as compared to its occurrence across the dataset, this will mean that this is an important predictor for authenticity. This also determines the rate of transformation and also includes the portion that makes this part unique and highly sophisticated. For the current dataset, it can be seen that this was done both to the text and the title. This was done in order to manipulate the importance of each word when it occurs in the title, along with the occurrence of each word with reference to the overall context of the text that was involved, it predicts with greater approach, that the words that are predominant to this essay are they important to the others, or in the determination of producing news that is more authentic than the one that is not authentic in nature, or is simply another feat for marketing.

3.6 Feature Engineering

After the determination of the words, their presences and their relative importance along with the chains that are of importance, there were certain factors that were left unaddressed. For example, as discussed above, punctuation is one of the key components, and length is the other, so how are these variables treated. Moreover, these keys or basic factors are also discussing.

Feature Creation

As stated above, the feature creation indicates the potential of each feature, since feature for every word is the first priority in NLP, therefore, it is recommended to that first, after the creation of the word features, there are two additional features that would be added besides the N-Grams, these are the length of the content and the percentage of prepositions that makes it necessary. This feature tends to outweigh the remaining of the factors that are important for discussion. The length of both the title and the percentage of prepositions are both very important in this dataset. Moreover, there was another variable termed as the source. This indicated the source or the origin point of the article; in other words, the website from where the article was taken. In creating the variables, and the overall analysis, this also led to the results and creation of dummies for each website. This made some extra binary features that indicated the source from where the article was chosen. This is not a bag of words, this has the real deal, so in the essence of the code, there would only be one source per row. This would make the test of which source is important in the determination of whether the article is factual, or without necessary facts to support itself.

Feature Combination

As it could be observed that the features that were combined were to be illustrated as the factor that makes the optimal combination of features, for this purpose, the optimal number of features was selected, and there was comprised of the features, and out of these features, a group of features were to be selected.

3.7 Model Building

The model that is being built has the sole purpose of determining the news and indicating whether this news is real or fake. Moreover, there are factors that should be considered, and in doing so, these factors should indicate whether the proportional and positional arguments make the model accurate. Here NLP is the model and the features that have been created are discussed above.

Model Selection

In order to select the model, the purpose is important, since this is a classification model. Therefore, the importance of this model lies in terms of the positional arguments. This is a typical decision tree, but to enhance the performance, an ensemble model of the random forest was applied to it, and this indicated that these factors tend to impact the model positively. Moreover, the variation of the model is based on a grid search, but in essence, this search is to find the best random forest model for analysis.

Viable Hyper Parameters

There are two major parameters that are of importance to a random forest model, and these two parameters indicate that the random forest model should be illustrative in terms of its number of features and the depth of the model. Moreover, the factors that matter most are the best settings and their combinations, so for the number of estimators the optimal choice is 10, 150 and 300, and for the depth, there are five choices, None, 30, 60, 90 and 120. It can be seen that these choices tend to illustrate the best results.

Grid Search of the model

For the purpose of the grid search, all possible combinations of the models, with three variations in the estimators, and 5 in the depths would be studied and compared, and these would be the metrics that would determine the potential variants in terms of analysis.

4.Results and discussion

The results are based on time for fitting the mode, time for getting results and the accuracy of the models.

4.1 Time for fitting the model

The first model is time for fitting the model, and it is divided into two parts, the mean of the model, and the standard deviation of the model.

Mean of Fitting models' time: In terms of the analysis of meantime, it can be said that with the decreased depth and the number of estimators, there is a lesser time to process each model, and it can be seen that the models with none as depth took more time than its counterparts.

Standard Deviation of Fitting Models' Time: The standard deviation is highest for models with 300 estimators and 60 depth. Moreover, it can be seen that the variation increases if there is 60 in-depth; it would increase the deviation.

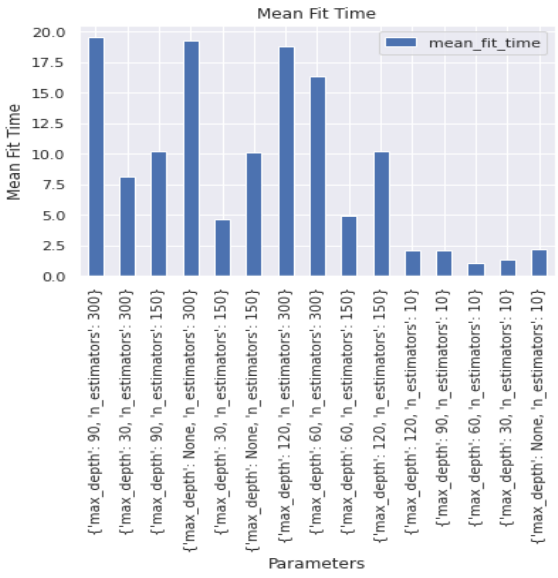

Figure 1. Mean of Fitting models' time

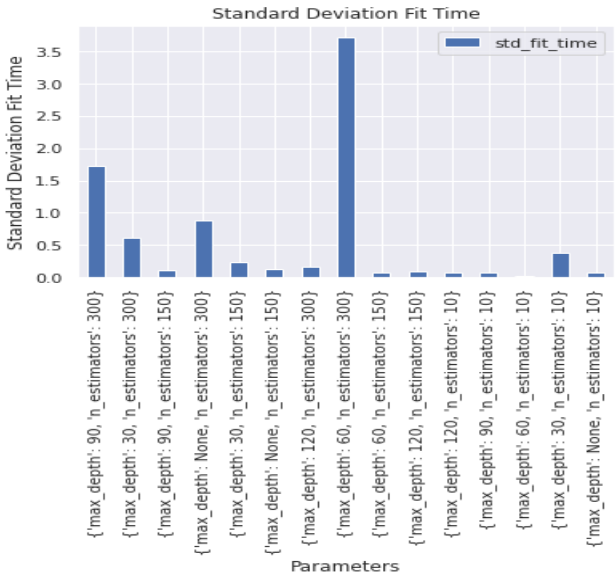

Figure 2. standard deviation of fitting Models' time

4.2 Time for Getting Results

Like the time for a fitting model, the time for results also has two parts

Mean Score Time: This is the scoring time of each model, and this indicates that there is a little change but when it comes to depth of 120 and 90, the scoring time increases, indicating that the trees of these widths are longer (Figure 3).

Standard Deviation of Score time: The scoring time of the models is similar to the standard deviation of fitting the models, and the results are then with the increase in both metrics, the results also increase (Figure 4)

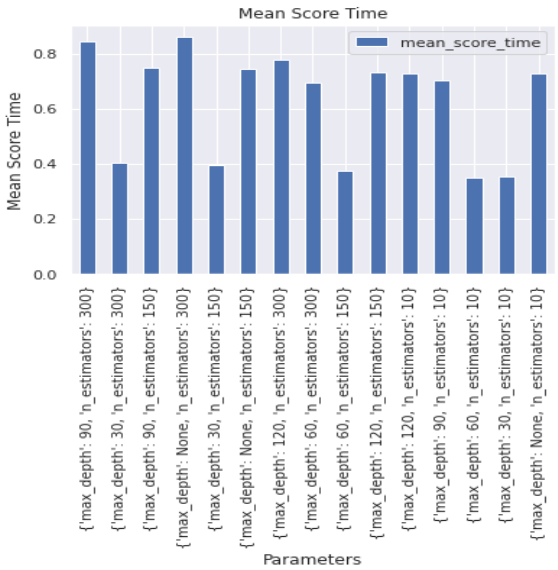

Figure 3. Mean Score Time

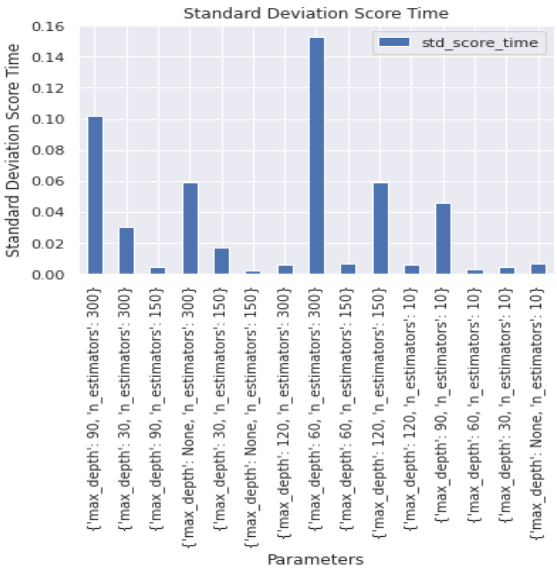

Figure 4. Standard Deviation of Score

4.3 Accuracy of the model

This has seven parts of accuracy, and 1 part of ranking on the basis of accuracy. These parts are the five variations of model and their respective accuracy, along with their mean and standard deviation.

Accuracy 1: This is the first test of accuracy, and it can be seen that the model parameters of 300 estimators with 60 depth are most accurate, and with the decrease in estimators, the accuracy also falls (Figure 5).

Accuracy 2: This is the second test of accuracy, and it can be seen that the model parameters of 300 estimators with 60 depth are most accurate, and with the decrease in estimators, the accuracy also falls (Figure 6).

Accuracy 3: This is the third test of accuracy, and it can be seen that the model parameters of 150 estimators with 90 depth are most accurate, and with the decrease in estimators, the accuracy also falls (Figure 7).

Accuracy 4: This is the fourth test of accuracy, and it can be seen that the model parameters of 300 estimators with 60 depth are most accurate, and with the decrease in estimators, the accuracy also falls (Figure 8).

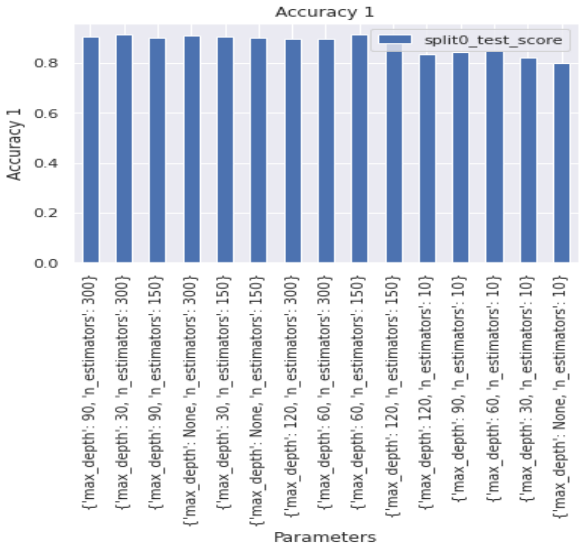

Figure 5. Accuracy 1

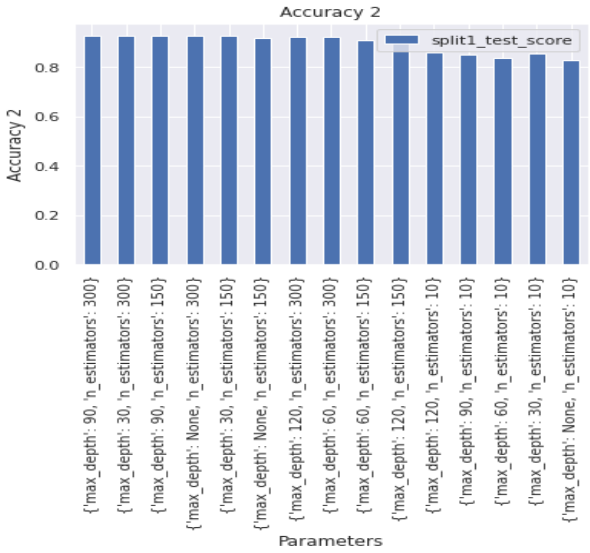

Figure 6. Accuracy 2

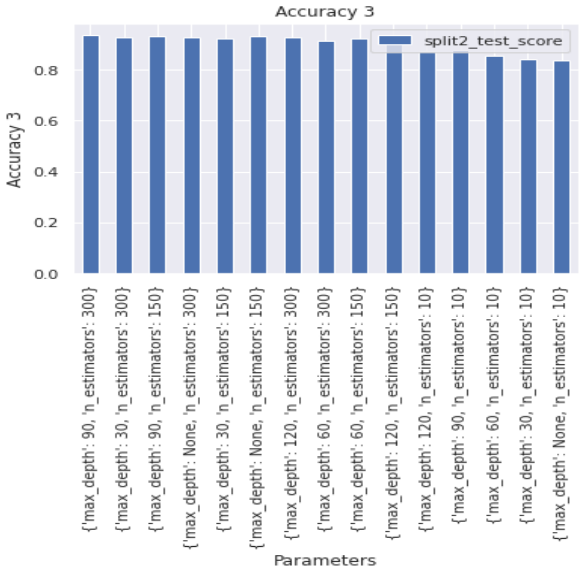

Figure 7. Accuracy 3

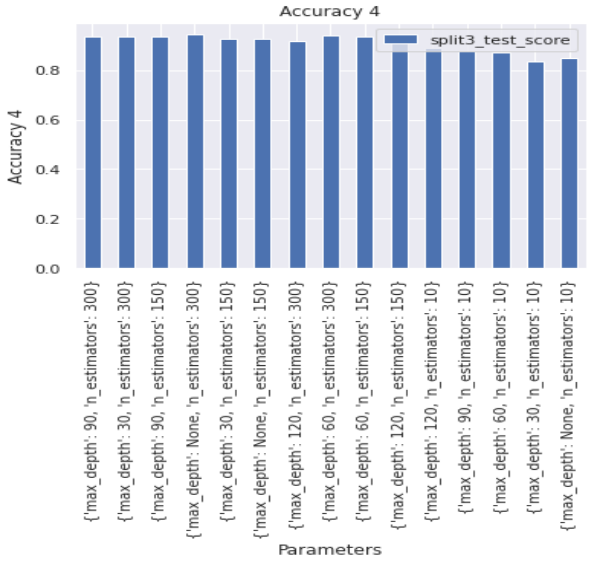

Figure 8. Accuracy 4

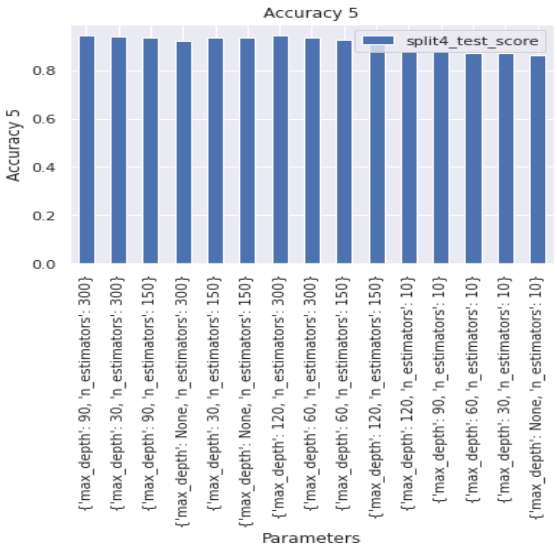

Figure 9. Accuracy 5

Accuracy 5: This is the fifth test of accuracy, and it can be seen that the model parameters of 300 estimators with 120 depth are most accurate, and with the decrease in estimators, the accuracy also falls. (Figure 9).

Mean Accuracy

This is the mean of all the values, and it can be observed that this means the results of max depth 90 and number of estimators to be 300. It also states that the accuracy is over 90% for this hurdle.

Standard Deviation of Accuracy

The standard deviation tells a different story, though there was a similarity in terms of the same variance and different heights indicate that the most consistent model was with 30 depth and 300 estimators followed by 60 depth and 150 estimators.

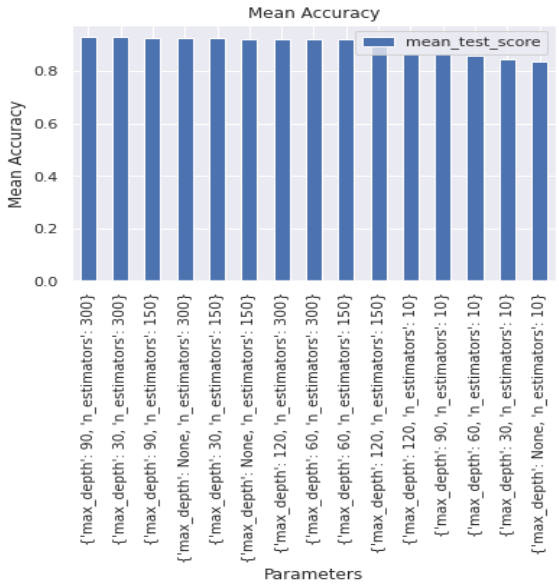

Figure 10. Mean Accuracy

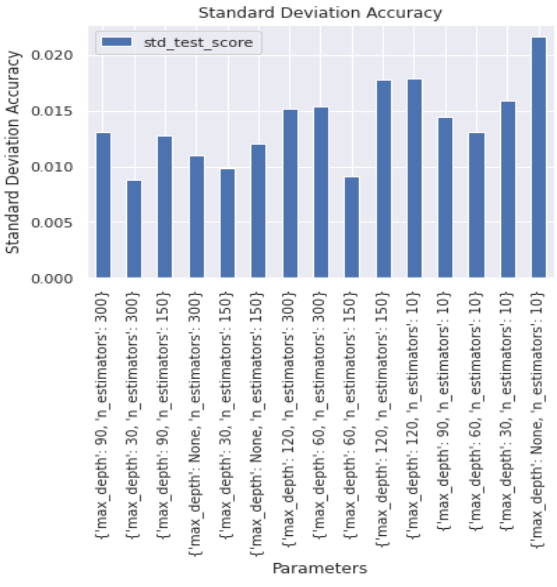

Figure 11. Standard Deviation of Accuracy

Ranking Means Accuracy

It can be observed that the ranking is given above though all the models are above 90%, the greatest variation is based on attaining a lower rank, leading to the best model, and here the depth of 90 with estimators of 300. And the worst model is with max depth 0 and 10 estimators. This indicates that more efficient the rank.

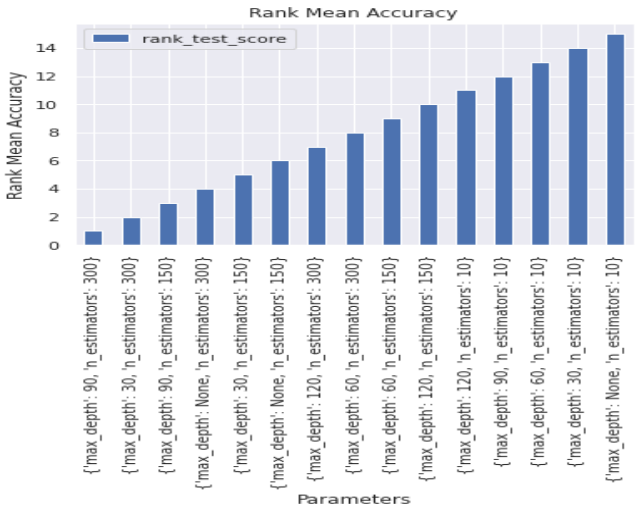

Figure 12. Ranking Means Accuracy

5.Conclusions

It can be concluded that in observing the overall study caused by the NLP model, and this model was conclusive on the basis of prediction, predicting more than 90 per cent and getting the results would indicate that the news or information can be identified in machine learning. It can therefore be said that in order to predict mechanically the periodic assessment tends to be the key, and this assessment indicates what the model is and what are its prediction based on arithmetic and numerical functions. This indicates that the model can predict with high accuracy, whether the news item on COVID-19. This algorithm can be improved upon, and it can be made evident that the algorithm is based on the numerical values that not only indicate the predictability but also enhance the forecasting speed and ability for the model to be ranked as genuine or simply misinformation. This can also help in collecting authentic articles and rejecting the ones that are useless, with considerable accuracy. The future scope might make an alert on the website, that this news has the probability of a certain figure to be false. Additionally, the objective can be attained, and the results forecasted are of great value to both individuals and organizations.

References

Yıldırım, S., Jothimani, D., Kavaklıoğlu, C., & Başar, A. (2018, December). Classification of" Hot News" for Financial Forecast Using NLP Techniques. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 4719-4722). IEEE

Alsudais, A., Tchalian, H., & Hilton, B. (2016). Labelled Topics for News Corpora Using Word Embeddings and Keyword Identification. In IJCAI Workshop on NLP Meets Journalism.

Li, Y., Zhang, J., & Yu, B. (2017, September). An NLP analysis of exaggerated claims in science news. In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism (pp. 106-111).

Bhatt, G., Sharma, A., Sharma, S., Nagpal, A., Raman, B., & Mittal, A. (2018, April). Combining neural, statistical and external features for fake news stance identification. In Companion Proceedings of The Web Conference 2018 (pp. 1353-1357).

Wankhede, K., & Shrawankar, U. (2016). Framing News Headline from Key Terms Using NLP. Proceedings of Advances in Engineering and Technology RICE, 16.

Dyson, L., & Golab, A. (2017). Fake News Detection Exploring the Application of NLP Methods to Machine Identification of Misleading News Sources. CAPP 30255 Adv. Mach. Learn—public Policy.

Jadhav, S. S., & Thepade, S. D. (2019). Fake news identification and classification using DSSM and improved recurrent neural network classifier. Applied Artificial Intelligence, 33(12), 1058-1068

Bhatt, G., Sharma, A., Sharma, S., Nagpal, A., Raman, B., & Mittal, A. (2017). On the benefit of combining neural, statistical and external features for fake news identification. arXiv preprint arXiv:1712.03935

Cokrowibowo, S., & Zulkarnaim, N. (2020, June). Online News Analysis of Majene Public Figure Electability with NLP (Natural Language Processing). In IOP Conference Series: Materials Science and Engineering (Vol. 875, No. 1, p. 012092). IOP Publishing.

Hamborg, F., Meuschke, N., Aizawa, A., & Gipp, B. (2017). Identification and analysis of media bias in news articles