Orchestrating and Interpreting Distributed Ensemble Learning for Enhanced Accuracy in Clickbait News Headline Detection

Monirul Haque
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

Sheikh Araf Noshin
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

Tanjina Akter Nipu
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

Shihab Uddin Sikder
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

Md Abu Ibrahim
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

A.M.Tayeful Islam
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

Sadiul Arefin Rafi
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

Adib Muhammad Amit
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

Annajiat Alim Rasel
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

***Abstract*—The rise of misinformation and misleading news has caused a massive outbreak on people doubting the reliability of the news sources. This research focuses on identifying if the news headline is a clickbait title or not to filter the message circulating throughout the internet media. Firstly, traditional statistical AI models such as Random Forest, Logistic regression, SVM, and Naive Bayes were run to set the benchmark for the accuracy. The novelty of this research is that from the voting system, three best-performing model was selected and used to run machine learning model in Ensemble learning to accurately predict the data. Two types of ensemble learning was used here : Bagging and Stacking. Both resulting in achieving accuracy of around 96% correct prediction among the 8000 test data consisting of mixed clickbait and not clickbait news headlines.**

***Index Terms*—Machine Learning, NLP, Binary Classification, Clickbait Classification, XAI, LIME, Gradient Boosting, Ensemble.**

I. INTRODUCTION

In this modern age of information technology and mass access of information granted by the vast revolution of the internet has its share of advantages and disadvantages as well. The accessibility of information may lead to misinterpretation and misleading information to be circulated throughout the world within a fraction of a second. This gave rise to the clickbait titles in the news/journal article that leads to misinformation and hindrance for the users. Clickbait is a term which refers to a scenario where an article/message/journal/video or any other form of data has misleading title/caption that influences the reader to think the news is about a specific topic of interest but actually upon reading the full document they are left clueless. This can also make a massive issue regardless since lots of the people only read the caption of the news and assume that this is the real news and this causes mass hysteria among the social media. Since due to the internet's availability, it's so easy to misunderstand and after a time when the news are rectified and shed light to the truth, the damage has been done. This research is focused on solving this issue by implementing different artificial intelligence models to classify and accurately predict whether the title is clickbait or not to enforce the reliability of the online media platforms and better monitoring to give the public reliable information. To accomplish that first a very reliable dataset was gathered and pre-processed accordingly. Different statistical models are run at first to understand the interpretability of the data and what can be the achievable goal for the prediction accuracy achieved. The purpose of this research is to provide an AI based solution which can be used in future to filter and monitor the news media to improve the reliability for the readers.

II. RELATED WORK

Clickbait's origins can be traced back to tabloids, a journalistic phenomenon that has existed since the 1980s [1]. The primary elements utilized for identifying clickbait are typically the attention-grabbing teaser sentence or text in a post, the article linked in the post that entices users to click, and the accompanying metadata [2]. Alongside the post text, which is commonly employed by most individ-

978-1-6654-5305-9/22/$31.00 ©2023 IEEE

uals to spot clickbait. The works of [3] and [4] connected article and the metadata were also considered and used handcrafted features, TF-IDF similarity between headline and article content and Gradient Boosted Decision Trees (GBDT). [2] suggested that clickbait detection should be a regression problem instead of a binary classification challenge, as the latter provides a way to measure how much clickbait is in the teaser message. They initiated the Webis clickbait challenge 2017, which boosted research activity in clickbait detection giving rise to highly effective and flexible deep learning techniques. [5] initially utilized self-attentive RNN [6] to choose the essential words in the title and then developed a BiGRU [7] network to encode the contextual information for the clickbait challenge 2017. On the other hand, [8] used an LSTM model [9] for the clickbait challenge that included article content. To construct the word embedding of clickbait titles, [10] employed the continuous skip-gram model [11]. However, [12] was the first to examine the use of transformer regression models in clickbait identification and finished first in the clickbait challenge.

III. Dataset

The dataset was compiled from Kashnitsky et el.[13] and Chakrobary et el.[14]'s dataset and then stratified with a perfect class balance of ratio of 50:50 with a total of 32000 text files containing headlines where 16000 texts are clickbait titles and the other 16000 are non-clickbait titles. The dataset is retrieved from various open-source online news and entertainment articles. Each value is a news article/message within that particular website that is labeled as either clickbait or non-clickbait. These headlines which we are defining as clickbait were collected from popular social news sites like 'ViralStories', Scoopwhoop', 'ViralNova', 'Thatscoop','Upworthy' and 'BuzzFeed'. These genuine and informative deadlines were gathered from reliable news sources like 'New York Times' , 'WikiNews' , 'The Hindu', and 'The Guardian'.

A. Annotation

The data was already labeled upon extraction. The dataset presents us with two columns: the first column comprises the headlines, while the second column contains numerical labels indicating the level of clickbait, where 1 signifies a clickbait headline and 0 signifies a non-clickbait headline. The dataset contains 32000 rows of which half of them are clickbait and the other half are non-clickbait titles. Furthermore, The dataset used for this research is a very well-balanced dataset with an approved inter-annotator agreement score in Fleiss' kappa score ranging around "moderate agreement" (Fleiss's kappa score 0.60). The value is not fixed since multiple resources and datasets are merged to make this combined dataset.

B. Class Breakdown and General Statistic

The data is equally distributed and both the clickbait and non-clickbait classes equal number of data. The dataset has 30340 numbers of words in it. Besides this, the dataset has 16074 total numbers of unique words in it. The number seems very big because the data was collected on different types of news headlines to make it more applicable to any type of news. On the clickbait data, the average text length is 55 words and on the non-clickbait data, the average text length is almost 50. This type of text has appeared more than 800 on clickbait data and more than 1000 on non-clickbait data.

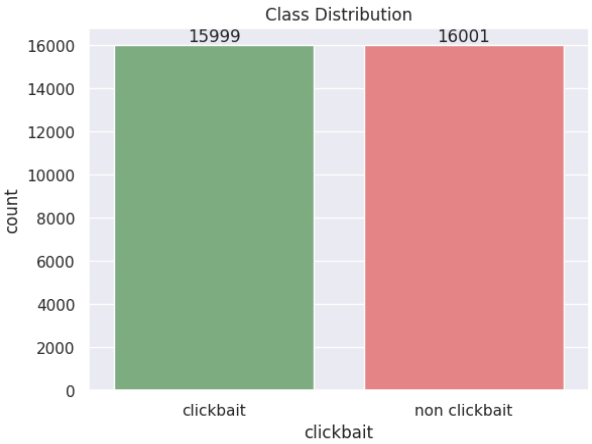

Figure 1: Class Breakdown

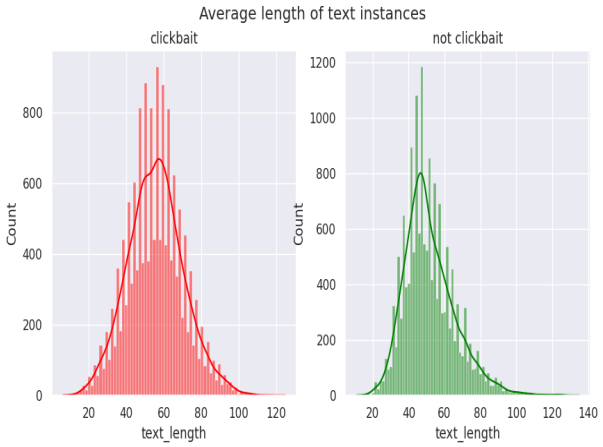

Figure 2: Text length size comparison

IV. Data Preprocessing

For our purpose, we have suggested a pre-processing method which is an essential initial step in classification and is fundamentally one of the essential steps in determining the recognition rate. This provides a way to reduce any sort of accuracy rate significantly and also ensures to help by normalizing the strokes and removing any irregularities. Data preprocessing is employed to address these concerns and transform them into decipherable data for the model. Below, we outline significant descriptions of some of these methods:

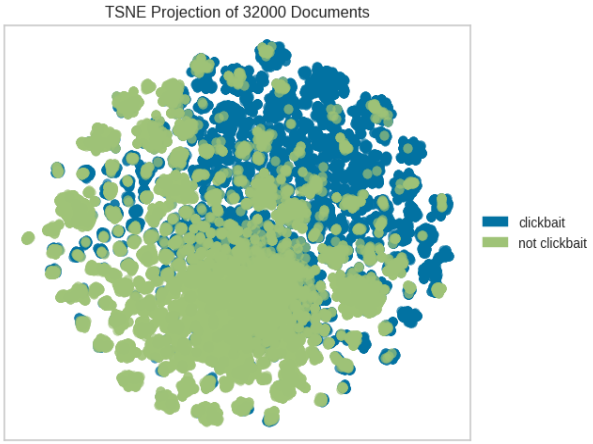

Figure 3: T-SNE projection

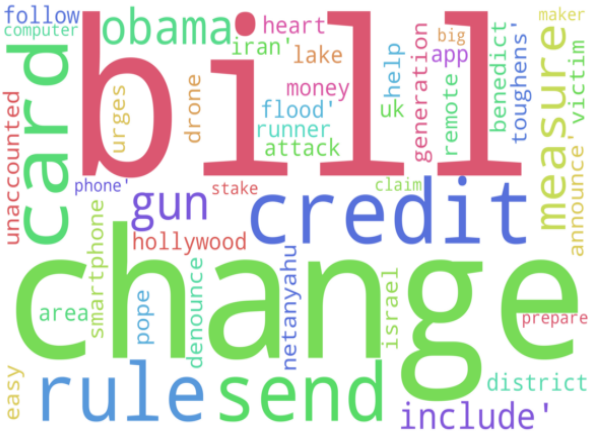

Figure 4: Wordcloud for Non-ClickBait

A. Stop word removal

For the cleanup of the data, firstly stop words were removed. i.e. “the”, “is” and “and”. Stop words refer to words that don’t usually hold any value in the understanding of the dataset for the AI models. Removal of these words will make the dataset more dense and help the AI to focus more on the important part to understand the context

B. Punctuation Removal

After removing stop words, other irrelevant characters such as : punctuation words, emojis, and emoticons were removed since they are redundant in the dataset. Removing all the unnecessary data and cleaning up the dataset will help the dataset to be more precise for the AI model to interpret.

C. Lemmatization

After the cleanup, Lemmatization was done to the text messages. Lemmatization is the process of grouping together different inflected forms of the same word.

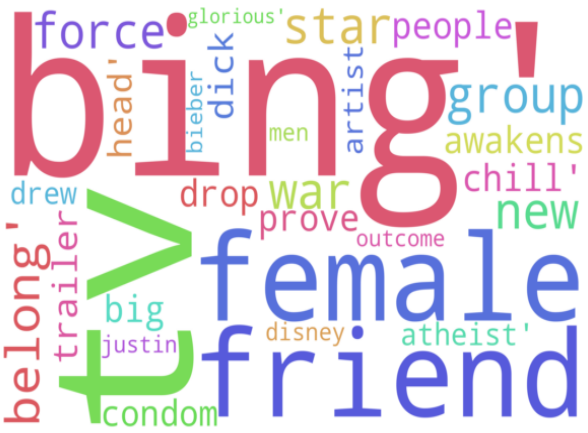

Figure 5: Wordcloud for ClickBait

Lemmatization connects words with similar meanings into a single word. Its objective is to simplify a word to its fundamental form, known as a lemma. For instance, the action ”running” would be recognized as ”run” through lemmatization.

D. TF-IDF vectorize

Finally, TF-IDF vectorizer was used to convert the text files into vectors. The TfidfVectorizer converts text into feature vectors that are suitable for use as input to the estimator. Essentially, it functions as a dictionary that translates every token (word) into a corresponding feature index within the matrix. Each distinct token is assigned a unique feature index.

V. Model Description

A lot of Models were used in this research. A brief detain about every model is given below:

A. Random forest

A powerful predictive model is produced using the ensemble learning technique known as Random Forest, which mixes various decision trees. It is named ”random” and ”forest” because it consists of a collection of decision trees and brings randomness into the process of building each individual tree. A typical supervised learning technique called a decision tree starts with a fundamental inquiry and then poses a succession of inquiries to arrive at an answer.A variety of tasks, including classification, regression, and feature selection, have been successfully completed using the flexible and effective machine learning method known as Random Forest.

B. Naive bayes

Naive Bayes, a probabilistic machine learning algorithm, operates on the principles of Bayes’ theorem. Operating under the assumption that the presence or absence of a particular feature is independent of the presence or absence of any other feature, it is characterized as being

"naive" since it makes this assumption. Naive Bayes has been utilized successfully in numerous applications despite this oversimplifying assumption, and it frequently performs well in real-world situations. Especially used in the field of classification and regression problems

C. SVM

The Support Vector Machine, commonly known as SVM, is a popular supervised learning method utilized for both regression and classification tasks. Finding non-linear decision boundaries and managing high-dimensional data are two areas where it excels. Its foundation is the notion of locating a hyperplane that best divides the data points of several classes. SVM is a good choice because it gives a decent accuracy on higher dimensions, can handle both linear and non-linear classification and has great robustness against overfitting.

D. XGBoost

XGBoost stands out as a widely recognized AI model, offering a scalable and exceptionally precise implementation of gradient boosting. It sets new benchmarks for computational capabilities within boosted tree algorithms. With each tree attempting to rectify the errors of the preceding one by utilizing the gradient of the error as the goal outcome, XGBoost develops an ensemble of shallow decision trees in a sequential manner. The weighted average of all the tree forecasts is the final projection.

E. CatBoost

Gradient boosting on decision trees serves as the learning technique for the CatBoost machine learning model. It is predicated on the notion of assembling weak models (decision trees) in a sequential manner to produce a strong model that can reduce the loss function. Some of the features that make this model a suitable model for this task is that it can handle the categorical feature, it uses oblivious trees with the implementation of order boosting and finally, it provides some tools for model analysis.

F. Light GBM

The learning method used by the Light GBM machine learning model is gradient boosting on decision trees. It is based on the idea that weak models (decision trees) may be sequentially put together to create a strong model that can lower the loss function. It uses histogram-based algorithms to split the data and it uses leaf-wise growth instead of level-wise growth. To increase the performance it uses two novel techniques which are GOSS or EFB.

G. Logistic regression

A machine learning model known as logistic regression employs a logistic function to compute the probability of a binary outcome, considering one or more predictor variables. It is one type of classification method that may give discrete labels to the data (such as 0 or 1) based on a certain probability threshold. Logit transformation, maximum likelihood estimation, and continuous and categorical are some of the features of logistic regression models.

H. Proposed Model: Ensemble Learning

Ensemble learning serves as a comprehensive meta-strategy in machine learning, aiming to enhance predictive accuracy by amalgamating predictions from multiple models. In this paper, after running all the statistical AI models three satisfactory trained models were chosen. This study employed two categories of ensemble learning methods: stacking and bagging.

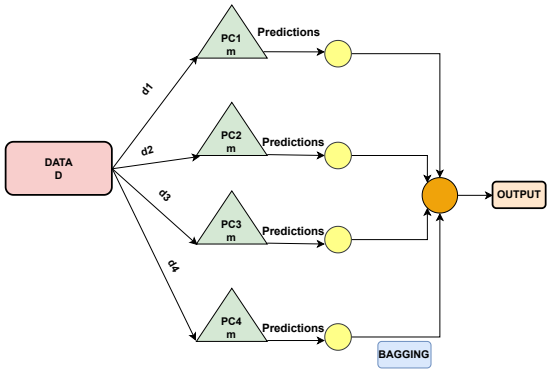

Figure 6: Ensemble Bagging in Distributed Systems

In the figure6 The first proposed architecture is given which is the ensemble learning technique Bagging incorporated with the distributed computing system environments. Bagging is a technique where the dataset or in this case, language corpus is split into different portions while maintaining the balance of the classes, and each equal portion is then fed into a different part of the total distributed system where each computer node has a decision tree based classifier that will parallel training chunk of data into each model. Then all the prediction result is accumulated and put through the aggregation process of max voting to get the final prediction. Here all the models are parallelly trained to achieve this goal.

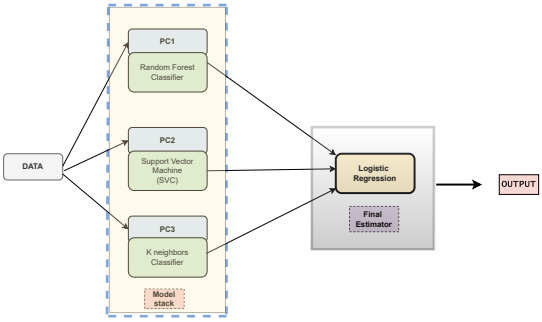

Figure 7: Stacking Ensemble in Distributed Systems

Additionally, the second proposed model for ensemble learning is Stacking. Stacking method is a bit different from the bagging. In this state, the data is not partitioned into many parts, rather the whole data is fed into different statistical classifier models to train their prediction and the resulting weighted parameters are then accumulated through some mathematical transformations and turned into a metamodel. Then the result is fed into a final estimator which is Logistic regression in this research and the output is generated. in the figure7 the proposed architecture is shown. The different computing system will act independently while training the statistical classifier and then the result will be fed into a central system with final estimator which will give the result.

I. XAI

This is an explainable artificial intelligence method which helps people to better understand how the AI models work. This is a very convenient way to interpret a black box AI model. The XAI technique applied in this research is referred to as LIME and SHAP, which stands for Local Interpretable Model-agnostic Explanations, is very popular XAI which makes a tabular format result to evaluate the model and how with each test data the prediction is shifting and the features that are responsible for it.

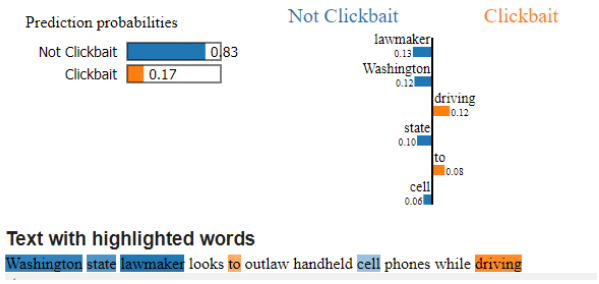

Figure 8: Explaination of a non-clickbait headline instance using lime

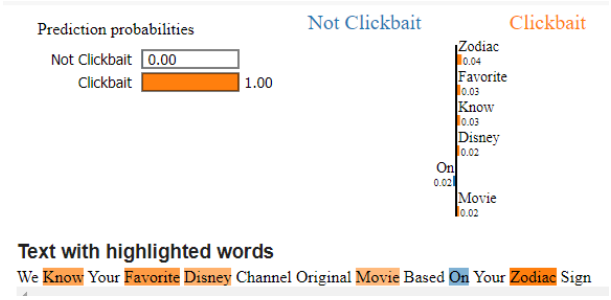

Figure 9: Explaination of a clickbait headline instance using lime

SHapley Additive exPlanations is a versatile machine learning interpretability tool that calculates fair and consistent feature importance values for explaining individual model predictions and helps to understand why a model made a specific prediction by attributing contributions of each feature.

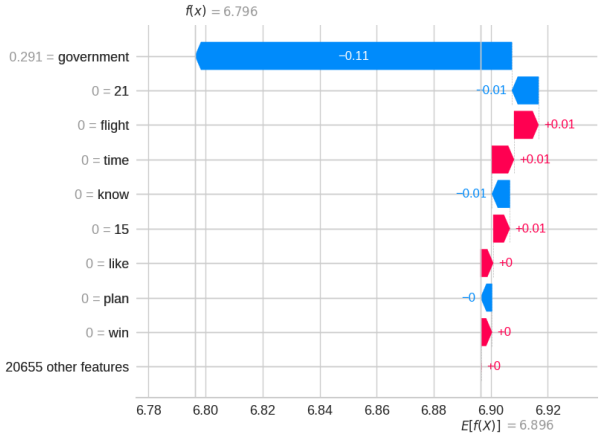

Figure 10: Explanation of a non-clickbait headline instance using SHAP

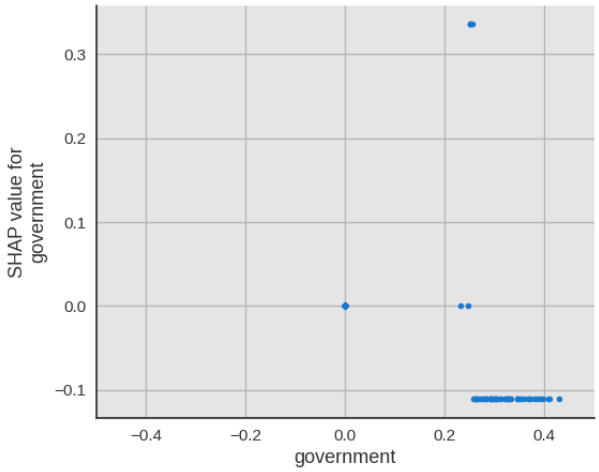

Figure 11: SHAP values for the word 'government' where positive refers to clickbait

VI. Result Analysis

Unseen data of 8000 were used to testify the proposed and pre-trained models. the train test split ratio is 75:25. At first the fundamental statistical AI models such as logistic regression, Random Forest, Naive Bayes Classifier, SVM were used. The results for respected models are given below. This helped to make a benchmark score which the research should attempt to overcome by running more complex model and ensemble model to unify these models and produce more accurate prediction to test the clickbait/non-clickbait titles. Also, for the sake of comparison, XGBoost, CatBoost, LightGBM models were run to compare and analyse the results.

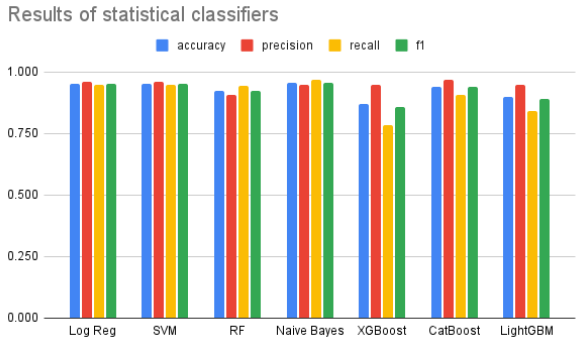

Figure 12: Testing Accuracy of Pre-trained Models

Figure 16 shows the Pre-trained models Testing accuracy. The better result-yielding models were SVM, Logistic Regression and Naive Bayes model all scoring well over 95% f1-score.

Table I: Summary Table of All Trained Models.

Model	Recall	Precision	F1 score	Accuracy
Naive Bayes	0.952	0.955	0.954	0.954
Logistic regression	0.948	0.960	0.954	0.954
Random forest	0.944	0.907	0.925	0.923
SVM	0.949	0.958	0.953	0.954
XGBoost	0.784	0.948	0.858	0.871
CatBoost	0.908	0.970	0.938	0.940
LightGBM	0.942	0.949	0.892	0.899
Bagging	0.966	0.949	0.958	0.956
Stacking	**0.967**	**0.948**	**0.958**	**0.957**

However, [15] precision is calculated by dividing the total number of predictions by the number of predictions that were made properly. The True Positive is calculated by dividing the True Positive and False Positive totals. Higher precision also suggests a better model, similar to recall score. The precision values show the same trend where the mentioned 3 models performed better.

Furthermore, the F1 score examines the harmonic mean of recall and precision, as explained in [15]. In plainer language, it unifies the evaluation of recall and precision into a single metric, enabling the comparison of several models. SVM, Logistic Regression, and Naive Bayes models all produced F1 scores more than 95% in the context of the F1 score.

The simplest statistic, accuracy [15], counts the proportion of accurate predictions among all predictions. This is not usually a valid metric of performance but in this case, since the dataset was balanced, accuracy also shows the results where the leading models are unchanged.

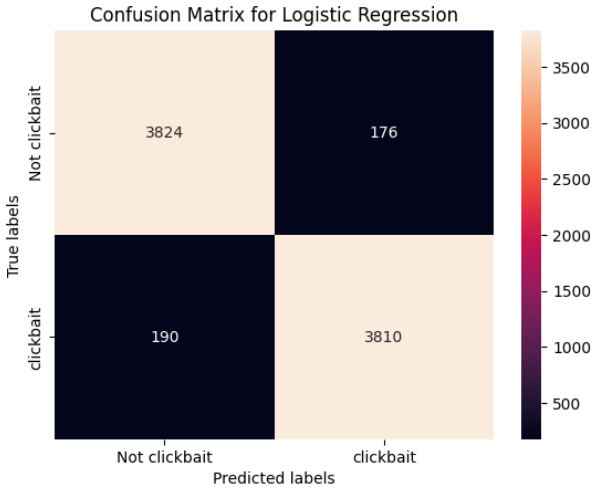

Figure 13: Confusion matrix of Stacking

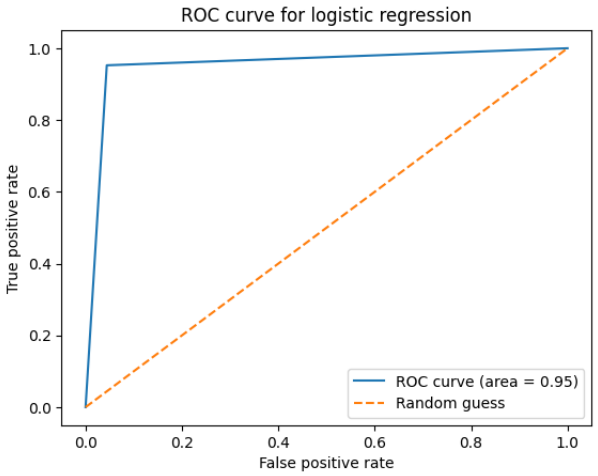

Figure 14: Roc of Bagging

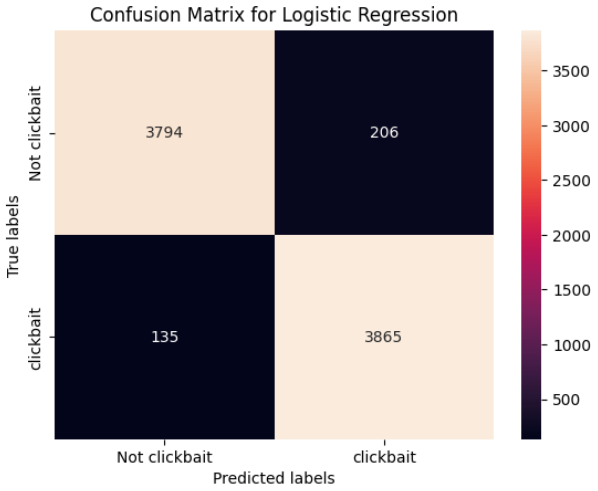

Figure 15: Confusion matrix of Bagging

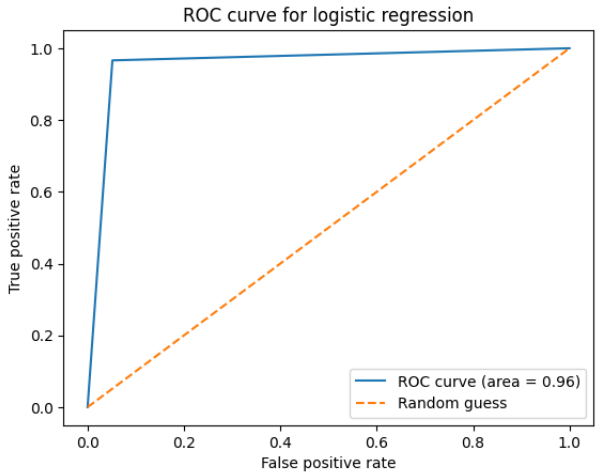

Figure 16: ROC of Stacking

Now our ensemble learning model's result is very apparent that it is superior to the traditional AI models. In the stacking ensemble model, the accuracy value was over 85%, and precision and recall also hovered around the same value. The f1-score is 95.41% which is above the standard. For the bagging method, the recall, precision, f1, and accuracy scores are 0.966, 0.949, 0.958 & 0.957 respectively.

Figure 10 showcases an example explanation from the SHAP model, where positive tends to be more clickbaity words. The word "government" is used in non-clickbait titles more than clickbait titles and so SHAP values for the word is more negative than positive. Figure 10 and 11 illustrate an example interpretation of the results XAI model LIME. here for each particular test data the results and affecting factors are shown through the illustrations of LIME model. In the above figure the orange part are responsible for the model to predict it as clickbait and the blue is for non-clickbait

VII. Conclusion

To conclude the research, this was a small effort to use power of AI models to pave a path in reliable systems or infrastructure that can help and monitor the data for varied news publications and help the concerned authorities to monitor and prevent misinformation spamming to stop. And the result reflects that ensemble learning is the way to make a viable solution for this issue. This research opens up a huge scope of future work to be done where more innovative data filtering and hyper-tuned model to get a more accurate result and make the monetizing of false media an achievable module for the media culture and for the consumers.

References

[1] S. E. Bird, "Tabloidization," *The International Encyclopedia of Communication*, 2008.

[2] M. Potthast, T. Gollub, M. Hagen, and B. Stein, "The clickbait challenge 2017: Towards a regression model for clickbait strength," *arXiv preprint arXiv:1812.10847*, 2018.

[3] M. Potthast, S. Köpsel, B. Stein, and M. Hagen, "Clickbait detection," in *European Conference on Information Retrieval*, Springer, 2016, pp. 810–817.

[4] P. Biyani, K. Tsioutsiouliklis, and J. Blackmer, "" 8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 2016. DOI: https://doi.org/10.1609/aaai.v30i1.9966.

[5] Y. Zhou, *Clickbait detection in tweets using self-attentive network*, 2017. arXiv: 1710.05364 [cs.CL].

[6] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[7] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, *On the properties of neural machine translation: Encoder-decoder approaches*, 2014. DOI: 10.48550/ARXIV.1409.1259. [Online]. Available: https://arxiv.org/abs/1409.1259.

[8] P. Thomas, *Clickbait identification using neural networks*, 2017. arXiv: 1710.08721 [cs.CL].

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[10] M. M. U. Rony, N. Hassan, and M. Yousuf, "Diving deep into clickbaits: Who use them to what extents in which topics with what effects?" In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, 2017, pp. 232–239.

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv:1310.4546*, 2013.

[12] V. Indurthi, B. Syed, M. Gupta, and V. Varma, "Predicting clickbait strength in online social media," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 4835–4846.

[13] Y. Kashnitsky, *Clickbait news detection*, 2019. [Online]. Available: https://kaggle.com/competitions/clickbait-news-detection.

[14] A. C. B. P. S. K. N. Ganguly, "Stop clickbait: Detecting and preventing clickbaits in online news media," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016. DOI: 10.1109/ASONAM.2016.7752207.

[15] E. Hussain, M. Hasan, S. Z. Hassan, T. H. Azmi, M. A. Rahman, and M. Z. Parvez, "Deep learning based binary classification for alzheimer's disease detection using brain mri images," *15th IEEE Con-*

ference on Industrial Electronics and Applications (ICIEA), pp. 1115–1120, 2020.