ACCURACY MEASURE FOR AUTOMATIC TOXIC SPEECH DETECTION USING NOVEL ADABOOST OVER RANDOM FOREST ALGORITHM

T.Varsha[1], K. Sashirekha[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: To compare and study the novel AdaBoost algorithm (NABA) and Random Forest algorithm for text wise toxic speech prediction for the purpose of enhanced accuracy of real-time voice detection.
Materials and Methods: The novel AdaBoost algorithm (N= 10) and Random Forest algorithm (N=10) methods are simulated by varying the NABA and random forest parameters to increase the pH. With the help of Gpower (80%) for two groups, the sample size is calculated as 20 samples per group for text analysis.
Results and Discussion: Based on obtained results NABA has significantly better accuracy (95.69%) compared to Random forest accuracy (80.33%). The statistical significance difference between AdaBoost and Random Forest was found to be $p=0.129$ ($p<0.05$) independent sample T-test value states that the groups are statistically insignificant.
Conclusion: AdaBoost algorithm produces better results in predicting toxic speech to improve accuracy percentage than the random forest algorithm.

Keywords: Speech Detection, Machine Learning, Novel AdaBoost, Convolutional Neural Network, Text Analytics, Feature Selection Algorithm

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. Saveetha University, Chennai, Tamil Nadu, India, Pincode:602105
[2*]Project Guide, Professor, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Science, Saveetha University, Chennai, Tamilnadu, India, Pincode 602105.

1. Introduction

In this research work, the Machine learning model had a great impact on Toxic speech prediction, which can predict values of present voice detection by working on the values that are already recorded (Alshalan et al. 2020). Feature selection algorithm for a toxic speech prediction is a trading platform where different people riot the internet society by their offensive speech (Ferri 2009). Prediction of toxic speech trends is taken into account as a crucial task and is of great attention as predicting toxic speech with supervised learning in success might result in engaging peace by making correct choices (Verma and Mohapatra, n.d.; Gonnet and Scholl, n.d.))(Verma and Mohapatra, n.d.; Gonnet and Scholl, n.d.)). toxic speech predictions are frequently used in social media for convolutional neural networks. (Gonnet and Scholl, n.d.) It has continually been a warm spot for online users and influencers to comprehend the alternate regularity of the internet usage by toxic speech and expect its trend in feature Selection Algorithm.Applications of hate speech detection are to reduce the toxicity in the social platform for inner peace and a friendly environment

In a research gap over the last 5 years, more than 65 papers have been published on ScienceDirect and google scholar on toxic speech predictions, which have major use cases in social media platforms and convolutional neural networks. A comparative gets text analytics predicting the speech can lead to profit in Feature Selection Algorithm (Nabipour et al. 2020). In this article (Mozafari, Farahbakhsh, and Crespi 2020) analysis of Random forest and NABA algorithm are implemented with the help of Applications for hate speech detection to reduce the toxicity in the social platform for inner peace and friendly environment experimental approach and also exhibited high-efficiency. This article (A. Ghosh et al., n.d.; S. Ghosh et al. 2021) presents the comparative analysis of the accuracy control of the novel AdaBoost Algorithm (NABA) using conventional controllers like speech prediction controllers and toxic speech detection (TSD) with text analytics. A novel method for NABA efficiency improvement using random forest algorithms and Toxic speed control is shown effectively. Our group has vast experience in various projects across numerous disciplines (Rivera and De Dios Santos Rivera 2020)

Our institution is keen on working on latest research trends and has extensive knowledge and research experience which resulted in quality publications (Rinesh et al. 2022; Sundararaman et al. 2022; Mohanavel et al. 2022; Ram et al. 2022; Dinesh Kumar et al. 2022; Vijayalakshmi et al. 2022; Sudhan et al. 2022; Kumar et al. 2022; Sathish et al. 2022; Mahesh et al. 2022; Yaashikaa et al. 2022). In a previous study the increase in efficiency of the text analytics SVM algorithm with toxic speech prediction was not properly considered to increase accuracy. The main aim of overcoming this issue is a novel Adaboost algorithm to improve the log loss rate of toxic speech prediction (Yin and Zubiaga 2021);(Reddy and Usha 2019).

2. Methods and Materials

This novel research work was carried out in the Machine Learning laboratory lab at Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. Using the G Power application, The sample size has been calculated by comparing both controllers in Supervised learning. The samples were separated into two groups for comparing the process and their result. In an individual group, 10 sets of samples and 20 samples in total are selected for this work. The pre-test power value is calculated using G-Power 3.1 software for convolutional neural networks (G power parameters: power=0.80, α=0.05). Two algorithms (NABA and random forest algorithm) are implemented using Technical Analysis software. In this research, no human and animal samples were used, so no ethical approval is required (A. Ghosh et al., n.d.)**.**

Adaboost Algorithm

Adaptive Boosting, or AdaBoost, majorly used as an Ensemble Method in Machine Learning as a boosting approach. Since the weights are re-allocated to each instance, with higher weights awarded to improperly identified examples, it is called Adaptive Boosting. It's used to boost the effectiveness of almost any machine learning model. It works well with students who are struggling. On a text analytics classification task, these are models that reach accuracy just above random chance. Decision trees with one level are the most suitable and hence most commonly used algorithm with AdaBoost. Both classification and regression issues can be solved with AdaBoost algorithms.

Random Forest Algorithm

One renowned supervised machine learning algorithm is RFA (Random forest algorithm). Random forest is used for both regression and classification which is given in equation (2). Random forest computation creates decision trees based on information testing, then gathers expectations from each one and determines the optimum arrangement by voting. It is a better option than a single choice tree since it reduces the complications of the result using equation (1). Random Forest algorithm pseudo-code.

$MSE = 1/N (fi - yi)^2$ —> equation (1)

Here, MSE = Mean squared error,
N = Number of data points recorded,
fi = The model's return value,
yi = The i data point's actual value,

Accuracy for AdaBoost and Random Forest algorithms was calculated based on the equation using Feature Selection Algorithm:

To calculate Accuracy, we use the formula TP + TN / TP +TN + FP + FN —> equation (2)
Here, TP = Total number of true positives,
TN = Total number of false negatives,
FN = Total number of true negatives,
FP = Total number of false positives,

The computer was equipped with an Intel Core i3 processor and 8GB of RAM. The system had a 64-bit operating system, an x64-based processor, and a 256-gigabyte SSD. The operating system is Windows 10, and the tool was Google Colab, which employed the Python programming language.

Statistical Analysis
In this research, statistical analysis of NABA and Random Forest algorithm-based methods are done using SPSS software. The independent variable is NABA accuracy, reduced toxic speech, and social norms, and the dependent variable is toxic comments, vulgar words, hatred, and mental health depression(Alshalan et al. 2020). The accuracy of the NABA is calculated using separate T-test analyses for both approaches.

3. Results

Table 1. shows the simulation result of the proposed algorithm AdaBoost and the existing system random forest which was executed at different intervals for a sample size of 10 using the google colab environment. From Table 1 it can be seen that the average accuracy of the NABA algorithm is 95.69% and the Random forest algorithm was 80.33%.
Table 2. represents the T-test comparison along with the Mean, Standard Deviation, and Standard Error Mean of both the NABA algorithm and the random forest algorithm. All the values among the study groups were calculated by taking an independent variable T-test using Feature Selection Algorithm. The LABA algorithm produces a significant difference from the RFA (random forest algorithm) with a P-value = 0.129 and effect size = 1.414.
Table 3. represents the Mean of the LSTM algorithm which is better compared with the random forest algorithm with a standard deviation of 0.22638 and 0.14062 respectively. From the results, the NABA algorithm (95.69%) gives better accuracy than the random forest algorithm (80.33%). Figure 1 gives the comparison chart of NABA of RFA (random forest algorithms) in terms of mean and accuracy with help of the Feature Selection Algorithm in a convolutional neural network. It can be observed that the NABA algorithm's mean accuracy is better than the Random forest. The error difference between the NABA algorithm (.07159) and the random forest algorithm (.04447) is shown in Figure 1.

4. Discussion

Consolidated results Based on obtained results NABA has significantly better accuracy (95.69%) compared to random forest accuracy (80.33%). Statistical significance difference between AdaBoost and Random Forest of single-tailed was found to be 0.129 ($p<0.05$) the Independent sample T-test value states that the results in the study are significantly not achieved. LABA and random forest algorithms are implemented and compared for toxic speech prediction to improve the accuracy of toxic speech. From the derived results, it is evident that the RFA (random forest algorithm) gives increased accuracy results compared to the LABA algorithm.

Proposed LABA algorithm for predicting toxic speech of selected social platforms by comparing the daily toxic speech movement in various sectors. (Raza 2017) implemented six machine learning techniques i.e., ANN, MLP, RBF, SVM, Decision Tree, and Naive Bayes and by comparing them concluded that MLP works better with an accuracy of 77%. Major research contribution supports Implementation and comparative analysis of random forest algorithms to optimize toxic speech of LABA drive with reduced efficiency improvement. Even though few articles listed the disadvantages of the proposed random forest algorithm. (Gagliardone et al. 2015) Furthermore, the random forest algorithm is not suitable for improving the accuracy of toxic speech prediction(Gupta et al., n.d.).

From the above discussion, only a few articles ensure that they provide better performance than the proposed LABA and random forest algorithm for improving the accuracy of toxic speech prediction. Also, there is no hidden or additional cost involved in the present price prediction model, so it received great support in the community in recent years (Cinelli et al. 2021). So, we can infer that the proposed LABA and random forest algorithm can be used to improve the accuracy of Speech prediction by regulating toxic speech (Gagliardone et al. 2015). Toxic speech prediction has limited speech prediction ability based on future text significant profit which makes better speech prediction in the future. Machine Learning algorithms can address future toxic speech prediction (Marinšek 2019).

5. Conclusion

The work involves the AdaBoost algorithm to find the toxic speech prediction with better proven accuracy of 95.69% in comparison to Random Forest accuracy is 80.33% for predicting toxic speech.

Declarations

Conflict of Interests

No conflict of interest in this manuscript.

Authors Contributions

Author VT was involved in data collection, data analysis and manuscript writing. Author KSR was involved in the conceptualization, data validation and critical review of manuscript.

Acknowledgements

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Cyclotron Technologies Pvt.Ltd, Chennai
2. Saveetha University
3. Saveetha Institute of Medical And Technical Science
4. Saveetha School of Engineering

6. References

Alshalan, Raghad, Hend Al-Khalifa, Duaa Alsaeed, Heyam Al-Baity, and Shahad Alshalan. 2020. "Detection of Hate Speech in COVID-19-Related Tweets in the Arab Region: Deep Learning and Topic Modeling Approach." *Journal of Medical Internet Research* 22 (12): e22609.

Cinelli, Matteo, Andraž Pelicon, Igor Mozetič, Walter Quattrociocchi, Petra Kralj Novak, and Fabiana Zollo. 2021. "Dynamics of Online Hate and Misinformation." *Scientific Reports* 11 (1): 22083.

Dinesh Kumar, M., V. Godvin Sharmila, Gopalakrishnan Kumar, Jeong-Hoon Park, Siham Yousuf Al-Qaradawi, and J. Rajesh Banu. 2022. "Surfactant Induced Microwave Disintegration for Enhanced Biohydrogen Production from Macroalgae Biomass: Thermodynamics and Energetics." *Bioresource Technology* 350 (April): 126904.

Ferri, Fred F. 2009. "Toxic Shock Syndrome." *Ferri's Color Atlas and Text of Clinical Medicine*. https://doi.org/10.1016/b978-1-4160-4919-7.50354-6.

Gagliardone, Iginio, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering Online Hate Speech*. UNESCO Publishing.

Ghosh, Achyut, Soumik Bose, Giridhar Maji, Narayan Debnath, and Soumya Sen. n.d. "Stock Price Prediction Using LSTM on Indian Share Market." https://doi.org/10.29007/qgcz.

Ghosh, Sreyan, Sonal Kumar, Samden Lepcha, and Suraj S. Jain. 2021. "Toxic Text Classification." *Data Science and Security*. https://doi.org/10.1007/978-981-15-5309-7_27.

Gonnet, Gaston H., and Ralf Scholl. n.d. "Stock Market Prediction." *Scientific Computation*. https://doi.org/10.1017/cbo9780511815027.008.

———. n.d. "Stock Market Prediction." *Scientific Computation*. https://doi.org/10.1017/cbo9780511815027.008.

Gupta, Archana, Pranay Bhatia, Kashyap Dave, and Pritesh Jain. n.d. "Stock Market Prediction Using Data Mining Techniques." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3370789.

Kumar, J. Aravind, J. Aravind Kumar, S. Sathish, T. Krithiga, T. R. Praveenkumar, S. Lokesh, D. Prabu, A. Annam Renita, P. Prakash, and M. Rajasimman. 2022. "A Comprehensive Review on Bio-Hydrogen Production from Brewery Industrial Wastewater and Its Treatment Methodologies." *Fuel*. https://doi.org/10.1016/j.fuel.2022.123594.

Mahesh, Narayanan, Srinivasan Balakumar, Uthaman Danya, Shanmugasundaram Shyamalagowri, Palanisamy Suresh Babu, Jeyaseelan Aravind, Murugesan Kamaraj, and Muthusamy Govarthanan. 2022. "A Review on Mitigation of Emerging Contaminants in an Aqueous Environment Using Microbial Bio-Machines as Sustainable Tools: Progress and Limitations." *Journal of Water Process Engineering*. https://doi.org/10.1016/j.jwpe.2022.102712.

Marinšek, Rok. 2019. *Cross-Lingual Embeddings for Hate Speech Detection in Comments: Master's Thesis : The 2nd Cycle Master's Study Programme Computer and Information Science*.

Mohanavel, Vinayagam, K. Ravi Kumar, T. Sathish, Palanivel Velmurugan, Alagar Karthick, M. Ravichandran, Saleh Alfarraj, Hesham S. Almoallim, Shanmugam Sureshkumar, and J. Isaac JoshuaRamesh Lalvani. 2022. "Investigation on Inorganic Salts K2TiF6 and KBF4 to Develop

Nanoparticles Based TiB2 Reinforcement Aluminium Composites." *Bioinorganic Chemistry and Applications* 2022 (January): 8559402.

Mozafari, Marzieh, Reza Farahbakhsh, and Noël Crespi. 2020. "Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model." *PloS One* 15 (8): e0237861.

Nabipour, M., P. Nayyeri, H. Jabani, A. Mosavi, E. Salwana, and Shahab S. 2020. "Deep Learning for Stock Market Prediction." *Entropy* 22 (8). https://doi.org/10.3390/e22080840.

Ram, G. Dinesh, G. Dinesh Ram, S. Praveen Kumar, T. Yuvaraj, Thanikanti Sudhakar Babu, and Karthik Balasubramanian. 2022. "Simulation and Investigation of MEMS Bilayer Solar Energy Harvester for Smart Wireless Sensor Applications." *Sustainable Energy Technologies and Assessments*. https://doi.org/10.1016/j.seta.2022.102102.

Raza, Kamran. 2017. "Prediction of Stock Market Performance by Using Machine Learning Techniques." *2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT)*. https://doi.org/10.1109/icieect.2017.7916583.

Reddy, Bhanuteja, and J. C. Usha. 2019. "Prediction of Stock Market Using Stochastic Neural Networks." *International Journal of Innovative Research in Computer Science & Technology*. https://doi.org/10.21276/ijircst.2019.7.5.1.

Rinesh, S., K. Maheswari, B. Arthi, P. Sherubha, A. Vijay, S. Sridhar, T. Rajendran, and Yosef Asrat Waji. 2022. "Investigations on Brain Tumor Classification Using Hybrid Machine Learning Algorithms." *Journal of Healthcare Engineering* 2022 (February): 2761847.

Rivera, Juan De Dios Santos, and Juan De Dios Santos Rivera. 2020. "Identifying Toxic Text from a Google Chrome Extension." *Practical TensorFlow.js*. https://doi.org/10.1007/978-1-4842-6273-3_6.

Sathish, T., V. Mohanavel, M. Arunkumar, K. Rajan, Manzoore Elahi M. Soudagar, M. A. Mujtaba, Saleh H. Salmen, Sami Al Obaid, H. Fayaz, and S. Sivakumar. 2022. "Utilization of Azadirachta Indica Biodiesel, Ethanol and Diesel Blends for Diesel Engine Applications with Engine Emission Profile." *Fuel*. https://doi.org/10.1016/j.fuel.2022.123798.

Sudhan, M. B., M. Sinthuja, S. Pravinth Raja, J. Amutharaj, G. Charlyn Pushpa Latha, S. Sheeba Rachel, T. Anitha, T. Rajendran, and Yosef Asrat Waji. 2022. "Segmentation and Classification of Glaucoma Using U-Net with Deep Learning Model." *Journal of Healthcare Engineering* 2022 (February): 1601354.

Sundararaman, Sathish, J. Aravind Kumar, Prabu Deivasigamani, and Yuvarajan Devarajan. 2022. "Emerging Pharma Residue Contaminants: Occurrence, Monitoring, Risk and Fate Assessment – A Challenge to Water Resource Management." *Science of The Total Environment*. https://doi.org/10.1016/j.scitotenv.2022.153897.

Verma, Nitish, and Baibaswata Mohapatra. n.d. "Stock Market Predication Using Machine Learning." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3645875.

Vijayalakshmi, V. J., Prakash Arumugam, A. Ananthi Christy, and R. Brindha. 2022. "Simultaneous Allocation of EV Charging Stations and Renewable Energy Sources: An Elite RERNN-m2MPA Approach." *International Journal of Energy Research*. https://doi.org/10.1002/er.7780.

Yaashikaa, P. R., P. Senthil Kumar, S. Jeevanantham, and R. Saravanan. 2022. "A Review on Bioremediation Approach for Heavy Metal Detoxification and Accumulation in Plants." *Environmental Pollution* 301 (May): 119035.

Yin, Wenjie, and Arkaitz Zubiaga. 2021. "Towards Generalisable Hate Speech Detection: A Review on Obstacles and Solutions." *PeerJ. Computer Science* 7 (June): e598.

Tables and Figures

Table 1. Predicted Accuracy of market (Accuracy of AdaBoost algorithm is 95.69% and Random forest algorithm is 80.33%)

S.No	Sample_size	AdaBoost algorithm accuracy in percentage	Random Forest algorithm accuracy in percentage
1	1	95.69	4.31

2	1	95.33	4.67
3	1	95.28	4.72
4	1	95.16	4.84
5	1	95.55	4.45
6	1	95.02	4.98
7	1	95.31	4.69
8	1	95.66	4.34
9	1	95.42	4.58
10	1	95.12	4.88

Table 2. Values of Mean, Standard deviation and standard error mean are obtained for 20 samples by statistical analysis of LABA and random forest algorithm..

	Algorithm	N	Mean	Std.deviation	Std.Error mean
Accuracy	Adaboost algorithm	10	95.3540	0.22638	0.07159
	Random forest	10	4.6420	0.14062	0.04447

Table 3. Independent sample T-test done for the two groups for significance and standard error determination. P>0.05 for wet basis.

		Levene's Test for Equality of Variances	T-test for Equality of means	
				95% confidence interval of the difference.

		F	Sig.	t	df	Sig.(2-tailed)	Mean Differences	Std.Error Differences	lower	upper
Accuracy	Equal Variances assumed	2.53	0.129	-.045	18	0.621	-0.040	0.8427	-.18105	0.173
		2.53	0.129	-.047	18	0.621	-0.040	0.8427	-.17305	0.181
	Equal Variances not assumed	2.53	0.129	-.047	18	0.621	-0.040	0.8427	-.17305	0.18

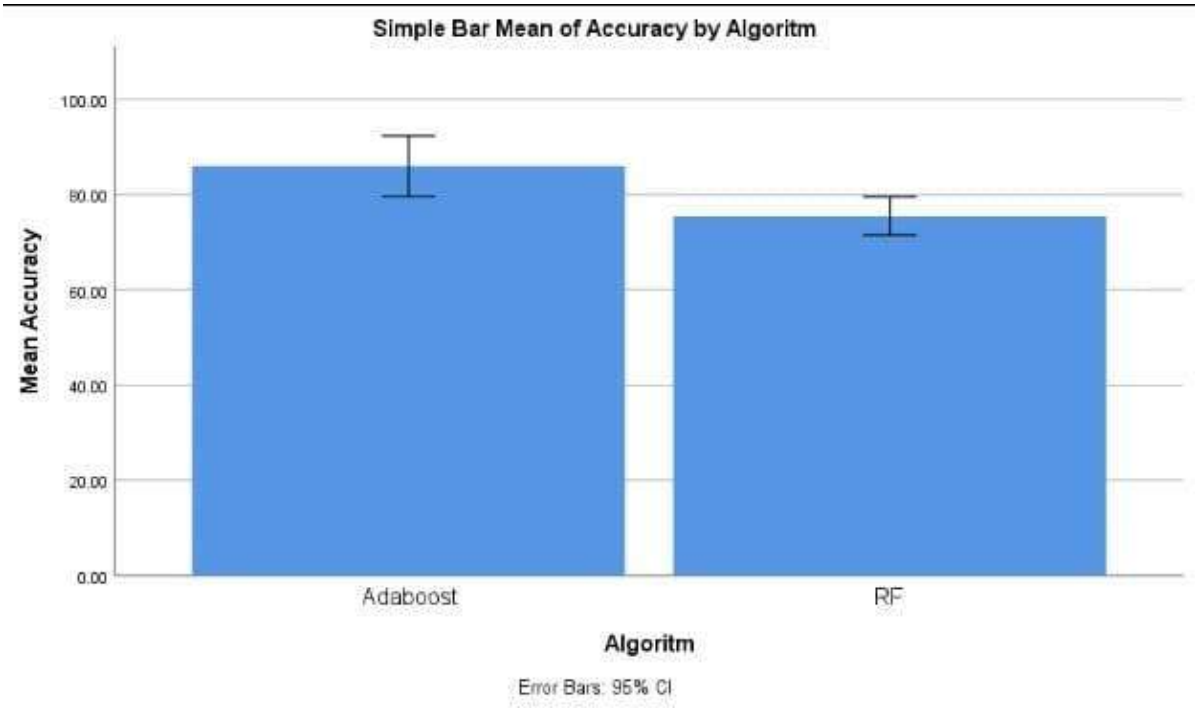

Fig. 1. Comparison of mean Accuracy of AdaBoost(95.69%) and Random forest (80.33%) model. The standard deviation appears to be less in the Random process when compared to the AdaBoost model. AdaBoost produces more consistent results than Random Forest. X- axis : Random forest and AdaBoost algorithms. Y-axis : Mean accuracy of Detection +/-1SD.