EFFICIENT PREDICTION OF TWITTER BOT IN SOCIAL MEDIA USING K- NEAREST NEIGHBORS COMPARED OVER LINEAR REGRESSION WITH IMPROVED ACCURACY

Kandala sridhar[1], S.Subbiah[2*]

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

Abstract:

Aim: Efficient prediction of twitter bot in social media using KNN compared over Linear Regression with improved accuracy. **Materials and Methods:** The KNN (N=10) and Linear Regression Algorithm (N=10) these two algorithms are calculated by using 2 Groups and I have taken 20 samples for both algorithm and accuracy in this work. **Results:** Based on the Results Accuracy obtained in terms of accuracy is identified by KNN(65.3%)over Linear Regression algorithm(75.9%).Statistical significance difference between KNN algorithm and Linear Regression Algorithm was found to be 0.220 ($p<0.05$). **Conclusion:**The Prediction of finding twitter bot in social media KNN when compared with Linear Regression.

Keywords: Twitter Bot In Social Media,Knn, Linear Regression, Classification Machine Learning.

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Science, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105 .

[2*]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

1. Introduction

Spam can be characterized as spontaneous undesirable, garbage email for a beneficiary or any email that the clients don't have any desire to have in their inboxes (Terrace 2009). Spam separating is a unique issue in the field of archive grouping and AI. In later years, the mechanical improvement in cell phones has expanded in computational power, and other strong frameworks have been proficient to be associated with cell phone organizations. This has likewise expanded the correspondence through SMS. No one needs the undesirable SMS on his PDA's back rub (Iliadis, Maglogiannis, and Plagianakos 2018). Furthermore, they need their inboxes to be liberated from such irritating SMS. It has specific characters that are unique in relation to messages. For example, email consists of specific organized data such as subject, mail header, greeting, source's location and so forth yet SMS needs such organized data (Iliadis, Maglogiannis, and Plagianakos 2018; Matloff 2017). Thus, to address SMS order issue, a similar strategy as applied to email can't be recreated here and have to foster a substance based separating approach which is much more troublesome. With regards to Nepali language, the issue isn't endeavored top to bottom and The dataset expected to explore isn't accessible (Matloff 2017). This examination work meant to make a SMS corpus and foster a structure to channel a Nepali Spam SMS. Since Nepali language has explicit trademarks (Etzion, Kuflik, and Motro 2006).

In Last 5 years 2017-2021 the Google Scholar has published more than 196 papers and the IEEE published more than 200 papers about twitter bot in social media. The analysis of KNN Algorithm and Linear Regression Algorithm in high performance efficiency has been made using an experimental approach. My study opinion is the efficient prediction of twitter bot in social media using a compressive of the twitter bot in social media prediction to Linear Regression.

The Accuracy of existing research is not properly existing in the system. The existence of the experiment is totally and the improvement of accuracy of a proposed algorithm system compared the existing model by improving (Etzion, Kuflik, and Motro 2006; Vetulani, Paroubek, and Innowacje 2019). To overcome these issues a KNN algorithm is implemented to improve twitter bot in social media in a network by comparing the proposed one with a Linear Regression Algorithm.Our team has extensive knowledge and research experience that has translated into high quality publications (Pandiyan et al. 2022; Yaashikaa, Devi, and Kumar 2022; Venu et al. 2022; Kumar et al. 2022; Nagaraju et al. 2022; Karpagam et al. 2022; Baraneedharan et al. 2022; Whangchai et al. 2022; Nagarajan et al. 2022; Deena et al. 2022) Now by the Above two Machine Algorithms that we have taken their own Advantages and Disadvantages in the Current survey[(Etzion, Kuflik, and Motro 2006; Vetulani, Paroubek, and Innowacje 2019; Saha, Kar, and Deb 2020)].On applying KNN Algorithm Memory to the Dataset followed by Performing Observations using Linear Regression and the results were plotted on a graph then there two techniques are(Pandiyan et al. 2022; Yaashikaa, Devi, and Kumar 2022; Venu et al. 2022; Kumar et al. 2022; Nagaraju et al. 2022; Karpagam et al. 2022; Baraneedharan et al. 2022; Whangchai et al. 2022; Nagarajan et al. 2022; Deena et al. 2022) compared based on the Result. Finally getting the best algorithm for predicting.

2. Materials and Methods

The research work is carried out in the Machine Learning laboratory lab at Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. The sample size has been calculated using the GPower software by comparing both of the controllers in Supervised learning . Two numbers of groups are selected for comparing the process and their result. In each group, 10 sets of samples and 20 samples in total are selected for this work. The pre-test power value is calculated using GPower 3.1 software (g power setting parameters: statistical test difference between two independent means, α=0.05, power=0.80, Two algorithms (E-Linear Regression and Linear Regression Algorithm) are implemented using Technical Analysis software. In this work, no human and animal samples were used so no ethical approval is required.

KNN Algorithm:

In measurements, the k-closest neighbors calculation is a non-parametric characterization technique initially created by Evelyn Fix and Joseph Hodges in 1951, and later extended by Thomas Cover. It is utilized for order and relapse. In the two cases, the info comprises of the k nearest preparing models in an informational collection.

Pseudocode KNN:

```
1.Calculate “d(x, xi)” i =1, 2, ….., n; where d denotes the Euclidean distance between the points.
2.Arrange the calculated n Euclidean distances in non-decreasing order.
3.Let k be a +ve integer, take the first k distances from this sorted list.
4.Find those k-points corresponding to these k-distances.
5.Let ki denotes the number of points belonging to the ith class among k points i.e. k ≥ 0
```

```
6.If $k_i > k_j \; \forall \; i \neq j$ then put x in class i.
```

Linear Regression Algorithm:

Linear Regression is a course of demonstrating the likelihood of a discrete result given an info variable. Calculated relapse is a valuable examination technique for arrangement issues, where you are attempting to decide whether another example squeezes best into a class.

Pseudocode Linear Regression:

```
1. For i←1 to k
2. For each training data instance di:
3. Set the target value for the regression to
        $Z_i \leftarrow \frac{Y_j - P(1|d_j)}{[P(1|d_j).(1-P(1|d_j))]}$
4. initialize the weight of instance dj to P (1| dj). (1-P(1| dj)
5. finalize a f(j) to the data with class value (zj) & weights (wj)
Classification Label Decision
6. Assign (class label:1) if P (1|dj) > 0.5, otherwise (class label: 2)
```

Statistical analysis

SPSS software is used for statistical analysis of novel approaches on efficient prediction of twitter bot in social media using KNN compared to Linear Regression with improved accuracy. The independent variable is KNN accuracy and the dependent variable is efficiency. The independent T test analyses are carried out to calculate the accuracy of the KNN for both methods.

3. Results

Below Table shows the simulation result of the proposed KNN algorithm and the existing system Linear Regression were run at different times in the google colab with a sample size of 10. From the table, it was observed that the mean accuracy of the Machine learning Algorithms like KNN was 80.91% and the Linear Regression algorithm was 69.88%.

The Mean, Standard Deviation and Standard Error Mean were calculated by taking an independent variable T test among the study groups[(Meghanathan, Chaki, and Nagamalai 2012)]. The KNN algorithm produces a significant difference than the Linear Regression algorithm with a value of 0.220 and effect size=1.612.

Table 2 represents the Mean of KNN algorithm which is better compared with the Linear Regression algorithm with a standard deviation of 0.71799 and 0.73395 respectively. From KNN algorithm and Linear Regression algorithm in terms of mean and accuracy[(Shamim Kaiser, n.d.)]. The mean results, the KNN(80.91%) gives better accuracy than the Linear Regression algorithm (69.88%). Figure 1 gives the comparison chart of KNN accuracy of the Linear Regression algorithm is better than Linear Regression. It is therefore, conclusive that KNN performs better than Linear Regression. The resultant plots are shown below in figure. The figure has been placed at the end of the paper[(Kravets et al. 2019)].

4. Discussion:

KNN and Linear Regression algorithms are implemented and compared for twitter bot in social media Prediction to improve the accuracy by review prediction[("Website," n.d.)]. From obtained results it is concluded that the KNN algorithm provides better accuracy results compared to the Linear Regression algorithm.

The principal concern of this study was to analyze the proficiency of Decision Trees, Back-proliferation Neural Organization and SVM based spam channels ("Website," n.d.; Corchado, Zunino, and Gastaldo 2008). The proficiency of these models were thought about based on generally utilized execution estimation measurements:- exactness, accuracy, review and F-score. Three AI calculations, to be specific, Decision Trees, Back-engendering Neural Network and Support Vector Machine were carried out with 446 highlights. These highlights were painstakingly chosen by completely investigating the idea of Nepali SMS ("Website," n.d.; Corchado, Zunino, and Gastaldo 2008; Singh et al. 2017). Observational examination of the exhibition of three Spam channels (Decision Trees, SVM and Multilayer Neural Network) for Nepali SMS dataset was done to view as the best one. It is seen from the trial that the Spam Filter based on Neural Network beat the Spam Filter in light of SVM also Decision Tree (Breiman 2017).

From the above discussion, only a few articles ensure that they provide better performance than the proposed KNN and Linear Regression algorithm for improving accuracy of twitter bot in social media prediction.So, we can infer that the proposed KNN and Linear Regression algorithm can be used to improve the accuracy (Breiman 2017; Hadi Amini et al. 2018).

5. Conclusion

Improved method for finding twitter bot in social media through ripening using KNN in comparison over Linear Regression with improved accuracy.The work involves KNN Prediction to be proved with better accuracy of 80.91% when compared to Linear Regression accuracy is 69.88%.

Declaration

Conflict of Interests

No conflict of interest in this manuscript.

Authors Contributions

Author SHV was involved in data collection, data analysis and manuscript writing. Author JJT 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. Soft Square Pvt.Ltd
2. Saveetha University
3. Saveetha Institute of Medical And Technical Sciences
4. Saveetha School of Engineering

6. References

Baraneedharan, P., Sethumathavan Vadivel, C. A. Anil, S. Beer Mohamed, and Saravanan Rajendran. 2022. “Advances in Preparation, Mechanism and Applications of Various Carbon Materials in Environmental Applications: A Review.” Chemosphere. https://doi.org/10.1016/j.chemosphere.2022.134596.

Breiman, Leo. 2017. Classification and Regression Trees. Routledge.

Corchado, Emilio, Rodolfo Zunino, and Paolo Gastaldo. 2008. Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems CISIS 2008. Springer Science & Business Media.

Deena, Santhana Raj, A. S. Vickram, S. Manikandan, R. Subbaiya, N. Karmegam, Balasubramani Ravindran, Soon Woong Chang, and Mukesh Kumar Awasthi. 2022. “Enhanced Biogas Production from Food Waste and Activated Sludge Using Advanced Techniques – A Review.” Bioresource Technology. https://doi.org/10.1016/j.biortech.2022.127234.

Etzion, Opher, Tsvi Kuflik, and Amihai Motro. 2006. Next Generation Information Technologies and Systems: 6th International Conference, NGITS 2006, Kebbutz Sehfayim, Israel, July 4-6, 2006, Proceedings. Springer.

Hadi Amini, M., Kianoosh G. Boroojeni, S. S. Iyengar, Panos M. Pardalos, Frede Blaabjerg, and Asad M. Madni. 2018. Sustainable Interdependent Networks: From Theory to Application. Springer.

Iliadis, Lazaros, Ilias Maglogiannis, and Vassilis Plagianakos. 2018. Artificial Intelligence Applications and Innovations: 14th IFIP WG 12.5 International Conference, AIAI 2018, Rhodes, Greece, May 25–27, 2018, Proceedings. Springer.

Karpagam, M., R. Beaulah Jeyavathana, Sathiya Kumar Chinnappan, K. V. Kanimozhi, and M. Sambath. 2022. “A Novel Face Recognition Model for Fighting against Human Trafficking in Surveillance Videos and Rescuing Victims.” Soft Computing. https://doi.org/10.1007/s00500-022-06931-1.

Kravets, Alla G., Peter P. Groumpos, Maxim Shcherbakov, and Marina Kultsova. 2019. Creativity in Intelligent Technologies and Data Science: Third Conference, CIT&DS 2019, Volgograd, Russia, September 16–19, 2019, Proceedings. Springer.

Kumar, P. Ganesh, P. Ganesh Kumar, Rajendran Prabakaran, D. Sakthivadivel, P. Somasundaram, V. S. Vigneswaran, and Sung Chul Kim. 2022. “Ultrasonication Time Optimization for Multi-Walled Carbon Nanotube Based Therminol-55 Nanofluid: An Experimental Investigation.” Journal of Thermal Analysis and Calorimetry. https://doi.org/10.1007/s10973-022-11298-4.

Matloff, Norman. 2017. Statistical Regression and Classification: From Linear Models to Machine Learning. CRC Press.

Meghanathan, Natarajan, Nabendu Chaki, and Dhinaharan Nagamalai. 2012. Advances in Computer Science and Information Technology. Computer Science and Information Technology: Second International Conference, CCSIT 2012, Bangalore, India, January 2-4, 2012. Proceedings, Part III. Springer.

Nagarajan, Karthik, Arul Rajagopalan, S. Angalaeswari, L. Natrayan, and Wubishet Degife Mammo. 2022. “Combined Economic Emission Dispatch of Microgrid with the Incorporation of Renewable Energy Sources Using Improved Mayfly Optimization Algorithm.” Computational Intelligence and Neuroscience 2022 (April): 6461690.

Nagaraju, V., B. R. Tapas Bapu, P. Bhuvaneswari, R. Anita, P. G. Kuppusamy, and S. Usha.

2022. "Role of Silicon Carbide Nanoparticle on Electromagnetic Interference Shielding Behavior of Carbon Fibre Epoxy Nanocomposites in 3-18GHz Frequency Bands." Silicon. https://doi.org/10.1007/s12633-022-01825-1.

Pandiyan, P., R. Sitharthan, S. Saravanan, Natarajan Prabaharan, M. Ramji Tiwari, T. Chinnadurai, T. Yuvaraj, and K. R. Devabalaji. 2022. "A Comprehensive Review of the Prospects for Rural Electrification Using Stand-Alone and Hybrid Energy Technologies." Sustainable Energy Technologies and Assessments. https://doi.org/10.1016/j.seta.2022.102155.

Saha, Ashim, Nirmalya Kar, and Suman Deb. 2020. Advances in Computational Intelligence, Security and Internet of Things: Second International Conference, ICCISIoT 2019, Agartala, India, December 13–14, 2019, Proceedings. Springer Nature.

Shamim Kaiser, M. n.d. Information and Communication Technology for Competitive Strategies (ICTCS 2020): Intelligent Strategies for ICT. Springer Nature.

Singh, Dharm, Balasubramanian Raman, Ashish Kumar Luhach, and Pawan Lingras. 2017. Advanced Informatics for Computing Research: First International Conference, ICAICR 2017, Jalandhar, India, March 17–18, 2017, Revised Selected Papers. Springer.

Terrace, Vincent. 2009. The Year in Television, 2008: A Catalog of New and Continuing Series, Miniseries, Specials and TV Movies. McFarland.

Venu, Harish, Ibham Veza, Lokesh Selvam, Prabhu Appavu, V. Dhana Raju, Lingesan Subramani, and Jayashri N. Nair. 2022. "Analysis of Particle Size Diameter (PSD), Mass Fraction Burnt (MFB) and Particulate Number (PN) Emissions in a Diesel Engine Powered by Diesel/biodiesel/n-Amyl Alcohol Blends." Energy. https://doi.org/10.1016/j.energy.2022.123806 .

Vetulani, Zygmunt, Partick Paroubek, and Wydawnictwo Nauka i. Innowacje. 2019. Human Language Technologies as a Challenge for Computer Science and Linguistics - 2019.

"Website." n.d. epal Telecome Authority(NTA), MIS report, 2017,[online] Available: http://nta.gov.np/en/mis-reports/ MIS_Poush_2074.pdf.

Whangchai, Niwooti, Daovieng Yaibouathong, Pattranan Junluthin, Deepanraj Balakrishnan, Yuwalee Unpaprom, Rameshprabu Ramaraj, and Tipsukhon Pimpimol. 2022. "Effect of Biogas Sludge Meal Supplement in Feed on Growth Performance Molting Period and Production Cost of Giant Freshwater Prawn Culture." Chemosphere 301 (August): 134638.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Advances in the Application of Immobilized Enzyme for the Remediation of Hazardous Pollutant: A Review." Chemosphere 299 (July): 134390.

Tables and Figures

Table 1. Accuracy Values for KNN and LR

S.NO	KNN	LR
1	87.100	84.900
2	87.00	83.20
3	87.50	84.60
4	86.10	84.00
5	85.40	83.90
6	84.00	81.20
7	83.30	83.00
8	84.10	84.70
9	83.00	83.50

10	85.40	84.40

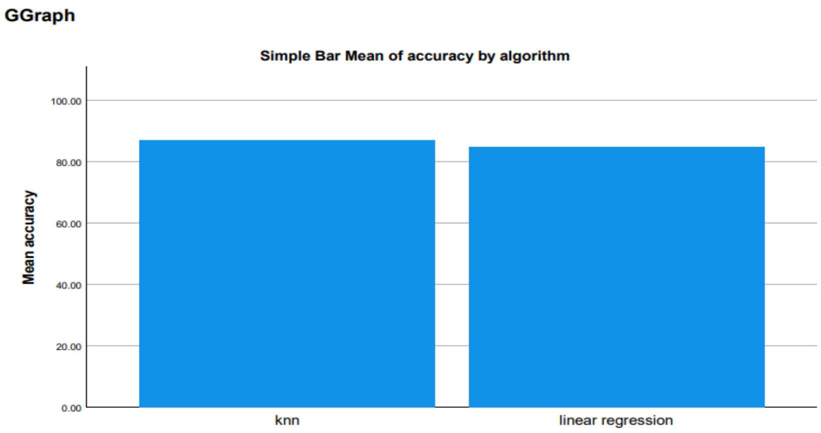
Table 2. Group Statistics Results-KNN has an mean accuracy (87.100%), std.deviation (3.09031), whereas for LR has mean accuracy (84.900%), std.deviation (2.07364).

Group Statistics					
Accuracy	**Groups**	**N**	**Mean**	**Std deviation**	**Std. Error Mean**
	CNN	10	87.1000	3.09031	1.38203
	LR	10	84.9000	2.07364	.92736

Table 3. Independent Samples T-test - KNN seems to be significantly better than LR (p=0.99

Accuracy	**Independent Samples Test**								
	Levene's Test for Equality of Variances					**T-test for Equality of Means**			
	F	**Sig**	**t**	**df**	**Sig(2-tailed)**	**Mean Difference**	**Std.Error Difference**	**95% Confidence Interval of the Difference**	
								Lower	**Upper**
Equal variances assumed	1.653	0.234	1.322	8	.223	2.20000	1.66433	-1.63796	6.03796
Equal variances not assumed			1.322	6.995	.228	2.20000	1.66433	-1.73610	6.13610

Fig. 1. Bar Graph Comparison on mean accuracy of KNN (87.10%) and LR (84.90%). X-axis: SVM, LR, Y-axis: Mean Accuracy with ±1 SD.


GGraph
Simple Bar Mean of accuracy by algorithm
Mean accuracy
knn
linear regression