BALTIC JOURNAL OF LAW & POLITICS
A Journal of Vytautas Magnus University
VOLUME 15, NUMBER 4 (2022)
ISSN 2029-0454

Cite: *Baltic Journal of Law & Politics* 15:4 (2022): 104-110
DOI: 10.2478/bjlp-2022-004010

Spam Detection on Emails Using Convolutional Neural Network Classifier with K nearest Neighbor Classifier

Yenimireddy Thirumala Kishon Reddy
Research Scholar,Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105

S. Sobitha Ahila
Project Guide, Corresponding Author, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105

Received: August 8, 2022; reviews: 2; accepted: November 29, 2022.

Abstract

Aim: The main aim of the research is to detect the spam emails using Convolutional Neural Network over KNN Algorithm and KNN algorithm belongs to Novel Cluster Based Method. **Materials and Methods:** Convolutional Neural Network and KNN are implemented in this research work. Sample size of n=20 is calculated using G power software. G power value is between 0.59 and 0.9 and determined as 10 per group with pretest power 80%, threshold 0.05% and CI 95%. **Result:** CNN algorithm provides a higher of 91.18% compared to KNN algorithm with 87.05% to classifise. There is a significant difference between two groups with a significance value of 0.003 ($p<0.05$). **Conclusion:** These results show that the performance of the Convolutional Neural Network algorithm (91.18%) detects spam emails better than KNN(87.05%) algorithm in terms of accuracy.

Keywords

Machine Learning,Spam Filtering,Convolutional Neural Network,Novel Cluster Based Method,K Nearest Neighbor,Black List,White List.

INTRODUCTION

The work is about Spam emails detection and spam filtering using Convolutional Neural Network over K Nearest Neighbor and KNN belongs to the Novel Cluster Based Method. The prediction using machine learning has succeeded in comparing Convolutional Neural Network over K Nearest Neighbor Algorithm. Nowadays email,text,and messenger have become part of our life. Spam email detection and email spam filtering is helpful to classify spam emails which saves a lot of time to the users which are using the emails and avoids frauds, theft of personal information from various hackers. Recently unwanted commercial emails are marked as spam emails (Mishra and Thakur 2013). Email spam is one of the major problem that is faced by the every email user. White list and black list plays a major role in classification of non-spam from spam emails. On a dialy basis every email user recieved hundreds of spam emails from various anonymous address. Some of the certain words are marked as the black list and white list. The normal methods for spam email filtering using black lists and white lists using ip address and email addresses (Laksono,

Basuki, and Bachtiar 2020). Applications of spam filtering were used in the classification of spam emails in our phones, laptops and in our offices.

In the last five years,Google Scholar identified almost 2000 articles on Spam email classification using Machine Learning. White list and black list plays a major role in classification of non spam from spam emails. Spam filtering in email detection plays a major role in the classification of normal emails from spam emails. Spam emails are very annoying to the email users who have fallen victim to the spam emails (Harisinghaney et al. 2014).Spam email classification is one of the most important tasks nowadays. Normally White list and Black list contain certain words that are used in spam and non-spam emails. It has been demonstrated that it is possible to use these machine learning algorithms to filter out spam emails from the normal emails (Sheshikala 2014). The purpose of these emails is to release confidiental personal information to the hackers and also to steal the passwords,usernames and Bank Verification Number (BVN),by using this information they try to steal money from our bank accounts (Cormack 2008; Sharma and Suryawanshi 2016). Some words are grouped into black list and white list to identify the emails as the spam or non spam emails and also spam filtering of emails. From all these research papers,the best study paper in my opinion is(Laksono, Basuki, and Bachtiar 2020).

Previously our team has a rich experience in working on various research projects across multiple disciplines (Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020; Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020).The research gap identified from the existing system shows poor accuracy. The study is to improve the accuracy of Classification by incorporating Convolutional Neural Network and comparing performance with K-Nearest Neighbour Algorithm. KNN is one of the parts in the Novel Cluster Based Method. The proposed model improves accuracy (4.13%) to achieve more efficiency (Tuteja and Bogiri 2016). The future scope of proposed work will be classification of spam emails based on using class labels for lesser time complexity.

MATERIALS AND METHODS

This study setting was done in the Soft Computing Laboratory,Saveetha School of Engineering,Saveetha Institute of Medical and Technical Sciences. Recall that the testing setup includes both hardware and software configuration choices. The laptop has an Intel Core i7 8th generation CPU with 12GB of RAM, an x86-based processor, a 64-bit operating system, and a hard drive. Currently,the software runs on Windows 10 and is programmed in Python. Once the program is finished, the accuracy value will appear. Procedure: Wi-Fi laptop connected. Chrome to Google Collaboratory search Write the code in Python. Run the code. To save the file,upload it to the disc,and create a folder for it. Log in using the ID from the message. Run the code to get the output accuracy and graph. The number of required samples in research are two in which group 1 is CNN compared with group 2 of K Nearest Neighbor Algorithm is part of the Novel Cluster Based Method. The samples were taken from the device and iterated 10 times to get desired accuracy with G power 80%, threshold 0.05% and CI 95%. The study uses a dataset spam or not-spam dataset consisting of a collection of spam and non spam emails downloaded from kaggle (mukulkirti 2021).

Convolutional Neural Network

CNN is one of the main parts of Neural Networks. CNN algorithm is utilized to detect the spam emails from the normal emails. This CNN is an efficient recogization algorithm which is widely used in pattern recognition and image processing (Santoso 2019). It has many functions such as simple structure, fewer parameters and adaptability. So it has become a hot topic in voice analysis and image recognition (Douzi et al. 2020).

Pseudocode for Convolutional Neural Network

Step1: Import packages.
Step2: Create an input dataset.
Step3: Analyze the size of the taken input data.

Step4: Split the datasets for testing and training the dataset.
Step5: Apply Convolutional Neural Network
Step6: Predict the results.

K Nearest Neighbour

K Nearest Neighbor algorithm is a simple algorithm that stores all available cases and classifies new cases based on the similarity measure (Sharma and Suryawanshi 2016). The KNN algorithm can compete with the most accurate predections. Therefore we use KNN algorithms for applications that require more accuracy. The quality of the predection depends on the distance measure (Nayak, Jiwani, and Rajitha 2021).

Pseudocode for K Nearest Neighbor

Step1: Import packages.
Step2: Create an input dataset.
Step3: Analyze the size of the taken input data.
Step4: Split the datasets for testing and training the dataset.
Step5: Apply K Nearest Neighbor algorithm.
Step 6: Predict the results.

Statistical Analysis

SPSS is a software tool used for statistics analysis. The proposed system utilized 10 iterations for each group with predicted accuracy noted and analyzed. Independent samples t-test was done to obtain significance between two groups. Dependent variable is no.of white list words and independent variable is no.of black list words. The study uses a dataset spam or not-spam dataset consisting of a collection of spam and non spam emails downloaded from kaggle (mukulkirti 2021).

RESULTS

Table 1 shows the accuracy value of iteration of CNN and KNN. Table 2 represents the Group statistics results which depicts CNN with mean accuracy of 91.18%,and standard deviation is 2.63. KNN has a mean accuracy of 87.05% and standard deviation is 1.21. Proposed KNN algorithm provides better performance compared to the CNN algorithm. Table 3 shows the independent samples T-test value for KNN and CNN with Mean difference as 4.12,Std Error Difference as 0.91. Significance value is observed as 0.003 ($p<0.05$). Figure.1 shows the bar graph comparison of mean of accuracy on KNN and CNN algorithm. Mean accuracy of CNN algorithm is 91.18% and KNN algorithm is 87.05%.

DISCUSSION

In this study,classification of spam emails using CNN algorithm has significantly higher accuracy,approximately 91.18% in comparison to KNN (87.05%). CNN algorithm appears to produce more consistent results with minimal standard deviation.

The similar findings of the paper (Sharma and Suryawanshi 2016) had an accuracy of 84% with CNN which was used to classify the spam emails from normal emails. The proposed work of (Sharma and Suryawanshi 2016) reported CNN has 84% accuracy which is used to classify the spam emails. The work proposed by (Sharma and Suryawanshi 2016) shows the CNN has a better accuracy of 90%. CNN algorithm which is used in both traditional and modern methods (Sharma and Suryawanshi 2016) as per their research it opposes CNN has highest accuracy and KNN algorithm will get least accuracy compared to other machine learning techniques which ranges between 60% when compared to other machine learning algorithms will get more accuracy than this. By using CNN to classify the spam emails it will have key issues to pretend (Guia, Silva, and Bernardino 2019; Renuka, Karthika Renuka, and Hamsapriya 2010) in this paper shows CNN has the least accuracy of 78%. Increasing the dataset's value only tends to get desired accuracy. CNN algorithm performs better with a combination of other machine learning algorithms.

The limitation of this research is that it cannot give appropriate results for smaller data like below the size of 5. In this model it is not able to consider all given feature variable parameters for training. The future scope of proposed work will be classification of spam emails based on using class labels for lesser time complexity.

CONCLUSION

These results show that the performance of the Convolutional Neural Network in classification of spam emails (91.18%) is better than that of the KNN (87.05%) in terms of accuracy. The Convolutional Neural Network algorithm has more efficiency than the KNN algorithm. Convolutional Neural Network algorithm classifies spam emails better than that of KNN algorithm.

DECLARATION

Conflict of Interests

No conflict of interests in this manuscript

Authors Contribution

Author YTKR was involved in data collection,data analysis,and manuscript writing. Author SSA was involved in conceptualization, data validation,and critical review of manuscript.

Acknowledgement

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. VK Technologies, Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

Benin, S. R., S. Kannan, Renjin J. Bright, and A. Jacob Moses. 2020. "A Review on Mechanical Characterization of Polymer Matrix Composites & Its Effects Reinforced with Various Natural Fibres." *Materials Today: Proceedings* 33 (January): 798–805.

Cormack, Gordon V. 2008. *Email Spam Filtering: A Systematic Review*. Now Publishers Inc.

Douzi, Samira, IPSS, Faculty of Science, University Mohammed Rabat, Morocco, Feda A. AlShahwan, Mouad Lemoudden, and Bouabid El Ouahidi. 2020. "Hybrid Email Spam Detection Model Using Artificial Intelligence." *International Journal of Machine Learning and Computing*. https://doi.org/10.18178/ijmlc.2020.10.2.937.

Gudipaneni, Ravi Kumar, Mohammad Khursheed Alam, Santosh R. Patil, and Mohmed Isaqali Karobari. 2020. "Measurement of the Maximum Occlusal Bite Force and Its Relation to the Caries Spectrum of First Permanent Molars in Early Permanent Dentition." *The Journal of Clinical Pediatric Dentistry* 44 (6): 423–28.

Guia, Márcio, Rodrigo Silva, and Jorge Bernardino. 2019. "Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis." *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. https://doi.org/10.5220/0008364105250531.

Harisinghaney, Anirudh, Aman Dixit, Saurabh Gupta, and Anuja Arora. 2014. "Text and Image Based Spam Email Classification Using KNN, Naïve Bayes and Reverse DBSCAN Algorithm." *2014 International Conference on Reliability Optimization and Information Technology (ICROIT)*. https://doi.org/10.1109/icroit.2014.6798302.

Laksono, Eko, Achmad Basuki, and Fitra Bachtiar. 2020. "Optimization of K Value in KNN Algorithm for Spam and Ham Email Classification." *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*. https://doi.org/10.29207/resti.v4i2.1845.

Mishra, Rachana, and R. S. Thakur. 2013. "Analysis of Random Forest and Naïve Bayes for Spam Mail Using Feature Selection Catagorization." *International Journal of Computer Applications*. https://doi.org/10.5120/13844-1670.

mukulkirti. 2021. "Naive_bayes_Theorem." Kaggle. September 29, 2021. https://kaggle.com/mukulkirti/naive-bayes-theorem.

Nalini, Devarajan, Jayaraman Selvaraj, and Ganesan Senthil Kumar. 2020. "Herbal Nutraceuticals: Safe and Potent Therapeutics to Battle Tumor Hypoxia." *Journal of Cancer Research and Clinical Oncology* 146 (1): 1–18.

Nayak, Rakesh, Salim Amirali Jiwani, and B. Rajitha. 2021. "Spam Email Detection Using Machine Learning Algorithm." *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.03.147.

Reddy, Poornima, Jogikalmat Krithikadatta, Valarmathi Srinivasan, Sandhya Raghu, and Natanasabapathy Velumurugan. 2020. "Dental Caries Profile and Associated Risk Factors Among Adolescent School Children in an Urban South-Indian City." *Oral Health & Preventive Dentistry* 18 (1): 379–86.

Renuka, D. Karthika, D. Karthika Renuka, and T. Hamsapriya. 2010. "Email Classification for Spam Detection Using Word Stemming." *International Journal of Computer Applications*. https://doi.org/10.5120/125-241.

Santoso, Budi. 2019. "An Analysis of Spam Email Detection Performance Assessment Using Machine Learning." *Jurnal Online Informatika*. https://doi.org/10.15575/join.v4i1.298.

Sathish, T., and S. Karthick. 2020. "Gravity Die Casting Based Analysis of Aluminum Alloy with AC4B Nano-Composite." *Materials Today: Proceedings* 33 (January): 2555–58.

Sathish, T., D. Bala Subramanian, R. Saravanan, and V. Dhinakaran. 2020. "Experimental Investigation of Temperature Variation on Flat Plate Collector by Using Silicon Carbide as a Nanofluid." In *PROCEEDINGS OF INTERNATIONAL CONFERENCE ON RECENT TRENDS IN MECHANICAL AND MATERIALS ENGINEERING: ICRTMME 2019*. AIP Publishing. https://doi.org/10.1063/5.0024965.

Sharma, Ajay, and Anil Suryawanshi. 2016. "A Novel Method for Detecting Spam Email Using KNN Classification with Spearman Correlation as Distance Measure." *International Journal of Computer Applications*. https://doi.org/10.5120/ijca2016908471.

Sheshikala, M. 2014. "IMPROVING SPAM EMAIL FILTERING EFFICIENCY USING BAYESIAN BACKWARD APPROACH PROJECT." *International Journal of Computer Science and Informatics*. https://doi.org/10.47893/ijcsi.2014.1164.

Sivasamy, Ramesh, Potu Venugopal, and Rodrigo Espinoza-González. 2020. "Structure, Electronic Structure, Optical and Magnetic Studies of Double Perovskite Gd2MnFeO6 Nanoparticles: First Principle and Experimental Studies." *Materials Today Communications* 25 (December): 101603.

Tuteja, Simranjit Kaur, and Nagaraju Bogiri. 2016. "Email Spam Filtering Using BPNN Classification Algorithm." *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*. https://doi.org/10.1109/icacdot.2016.7877720.

Venu, Harish, and Prabhu Appavu. 2021. "Experimental Studies on the Influence of Zirconium Nanoparticle on Biodiesel–diesel Fuel Blend in CI Engine." *International Journal of Ambient Energy* 42 (14): 1588–94.

TABLES AND FIGURES

Table 1. Accuracy Values for CNN and KNN

S.NO	CNN	KNN
1	89.76	86.12
2	89.89	85.77
3	87.05	86.95
4	87.45	88.89
5	93.15	87.59
6	93.45	86.75
7	90.28	85.33
8	94.07	86.45
9	93.58	87.95
10	93.17	88.76

Table 2. Group statistics results state that CNN has an mean accuracy (91.18%), std.deviation (2.63), whereas KNN has mean accuracy (87.05%), std.deviation (1.21).

Group Statistics					
	Groups	**N**	**Mean**	**Std deviation**	**Std. Error Mean**
Accuracy	CNN	10	91.18	2.63	0.83
	KNN	10	87.05	1.21	0.38

Table 3. Independent samples T-test - CNN seems to be significantly better than KNN (p=0.003)

	Independent Samples Test							
	Levene's Test for Equality of Variances					**T-test for Equality of Means**		
Accuracy	**F**	**Sig**	**t**	**df**	**Sig (2-tailed)**	**Mean Difference**	**Std.Error Difference**	**95% Confidence Interval of the Difference**

								Lower	Upper
Equal variances assumed	11.643	0.3	4.496	18	0.003	4.12	0.91832	2.19967	6.05833
Equal variances not assumed			4.496	12.669	0.003	4.12	0.91832	2.13980	6.11820

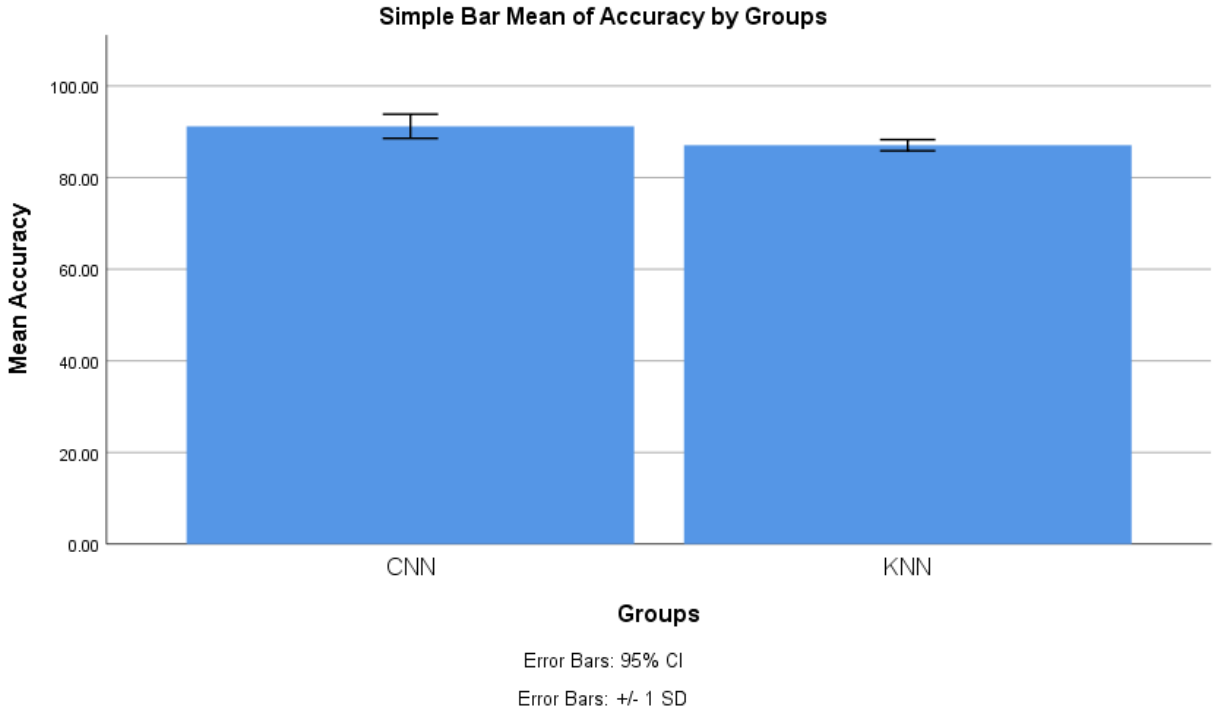

Fig.1.Bar Graph Comparison on mean accuracy of CNN (91.18%) and KNN (87.05%). X-axis:KNN,CNN algorithms, Y-axis: Mean Accuracy with ±1 SD.