Using the K-Nearest Neighbors Algorithm and Logistic Regression to Improve Accuracy, a Novel Machine Learning Approach for Detecting SMS Spam Message

B.Nithin Sai[1], Dr. B. Swaminathan[2*]

[1]Research Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Tamil Nadu, India, Pincode:602105

[2*]Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Science, Saveetha University, Tamil Nadu, India, Pincode: 602105

ABSTRACT

Aim: A Novel Approach for detecting SMS Spam effective model using Logistic Regression Algorithm to improve the accuracy and comparing with K-Nearest Neighbors Algorithm. **Materials and Methods:** By Using Classification of K-nearest Neighbors Algorithm (N=10) and Logistic Regression Algorithm (N=10) is performed in an algorithm. **Discussion:** This paper examines spam messages and offers their outline and outcomes. SMS detection may be regarded as a critical aspect in establishing and maintaining an algorithm.**Results:** In the detection of the Effective model of SMS spam messages with a Logistic Regression Algorithm accuracy is 96.0% compared KNN algorithm accuracy is 89.0%. The two Algorithms Logistic Regression Algorithm and KNN are statistically significant with the independent sample T-Test value is 0.04 ($p<0.05$) with a confidence level of 96% accuracy. **Conclusion:** Within the limit of the study analyzing the Effective model of spam using K-Nearest Neighbors Algorithm (KNN) and LR detected by SPSS tool and it was observed that the LR Algorithm significantly seems to be better than KNN Algorithm.

Keywords: Logistic Regression, Novel Spam Detection, K-nearest Neighbors, Spam filtering process, SMS, Spam messages

INTRODUCTION

SMS is a text messaging service that allows users to send and receive spam messages. Commercial adverts may also be sent to users' mobile phones via text messages (Akbari and Sajedi 2015). The amount of spam varies from one place to the next. There is a significant difference between spam-filtering in text messages and spam-filtering in emails. A substantial dataset is available for email, whereas the dataset for SMS Spam detection is small (Sasaki and Shinnou 2005). SMS is a text messaging service that allows phone users to send and receive short text messages. Because the text is shorter than email, the number of features needed to classify it is also smaller. The majority of text messages are short and use less formal terminology (Nagwani and Sharaff 2017). SMS spam detection is complicated by a number of factors, including a low rate of SMS spam, which has led to many users and service providers ignoring the problem (Shakiba, Zarifzadeh,

and Derhami 2018). Spam-filtering software is extremely limited. SMS spam filtering on the recipient's smartphone isn't a perfect solution. Spam is a term used to describe wanted and unwanted electronic messages (Shakiba, Zarifzadeh, and Derhami 2018; Saeedian and Beigy 2009). Spammers send these messages for a variety of reasons. This is done in order to obtain personal information from users (Liu, Lu, and Nayak 2021). Supervised learning is not the same in every place, and it can differ. The phone number is limited to sending 200 messages per hour and 1,000 Spam messages per day. SMS spammers have altered these methods in novel ways as a result. In order to accurately filter SMS spam detection, more effective ways are required (Alzahrani and Rawat 2019).

According to numerous recent spam messages reports, the vast majority of SMS on the internet ((Rajalingam 2020). According to a May 2009 Symantec report, 90.4 percent of SMS were spam. A Google report from September 2009 estimated spam volume to be 90-95 percent in all four quarters, while a Microsoft report from the same month estimated spam volume to be 97.3 percent of SMS (Akinyelu 2021). Botnets are thought to be responsible for roughly 85 percent of all spam, according to several reports. 56.7 percent of spam detection comes from recognized botnets, whereas a study claims that only six botnets are responsible for 79 percent of spam messages landing on the University of Washington campus (Rafique and Abulaish 2012).(Parakh et al. 2020; Pham et al. 2021; Perumal, Antony, and Muthuramalingam 2021; Sathiyamoorthi et al. 2021; Devarajan et al. 2021; Dhanraj and Rajeshkumar 2021; Uganya, Radhika, and Vijayaraj 2021; Tesfaye Jule et al. 2021; Nandhini, Ezhilarasan, and Rajeshkumar 2020; Kamath et al. 2020)

The research gap in the existing system is less efficient in SMS Spam detection systems in the spam filtering technique. Clustering automatically splits the dataset into groups based on their similarities. Anomaly detection can discover unusual data points in your dataset. Association mining identifies sets of items that often occur together in your dataset supervised machine learning finds all kinds of unknown patterns in data (Bosaeed, Katib, and Mehmood 2020). Naive Bayes Algorithm helps you to find features that can be useful for categorization. It takes place in real-time, so all the input data is to be analyzed and labeled in the presence of learners. It is easier to get unlabeled data from a computer than labeled data, which needs manual intervention (Ali and Maqsood 2018).

MATERIALS AND METHODS

This section explains the general framework of the study's work method. The dataset is analyzed and classified using an AI instrument in this investigation. At the most basic level, data is gathered from a variety of sources to create a useful dataset of ham and spam messages in text format, which is then used as the model's input. We converted the informational collection, which was previously in text format, to CSV at the second level of the study (Comma Separated Value).

The group1 (N=10) in the KNN Algorithm which is an existing system and group 2 (N=10) is thelogistic Regression algorithm with a sample size of 10 and the KNN algorithm is Group 2 with a sample size of 10, then they are compared for more accuracy and prediction score values for choosing the best. The data set used in this

study can be found on Kaggle, a machine learning repository. For better data quality, pre-processing is done, either by deleting unnecessary words or numbers. K-Nearest Neighbors (KNN) is a simple and fundamental machine learning method that is versatile and one of the best. KNN and Logistic Regression Algorithm is used here to compare the data. This includes accounting, medical services, political theory, image recognition, and video recognition. Attribution budgetary enterprises will anticipate the FICO assessment of clients in their evaluations. The KNN algorithm is used to solve classification and regression problems. The calculation of KNN is reliant on the similarity method (Karasoy and Ballı 2021).

The observations about spam on today's Internet suggest developing an Logistic Regression Algorithm scheme that automatically identifies the common terms shared by spam belonging to the same campaign, identifies the spam detection campaign and extracts the campaign signatures, to be used for filtering future spam of the same campaign. IT is theoretically feasible and highly promising. As explained in the spam filtering process, one of the main reasons that supervised learning requires labeled data is that it is difficult to identify spam from ham when looking at SMS one at a time. When the collective common qualities of spam detection, at the campaign level are studied together, a clear manifestation emerges. In a spam campaign, the scammer sends a huge number of spam messages with the same aim in an automated manner. For example, through a botnet. Legitimate SMS, on the other hand, are manually sent out by humans for various objectives. Such a basic difference is necessarily mirrored in the entropy of the SMS contents; Spam messages from a campaign have a low entropy, while ham has high entropy. By successfully detecting common unchanging text sections among the spam filtering process (Oecd and OECD 2011; Bishara, Bishara, and University 2016; Diale, Celik, and Van Der Walt 2019).

While spammers always try to enhance the entropy of spam by improving their obfuscation tactics, we claim that the entropy gap between spam or ham will never close due to the fundamental difference between campaign-based, legitimate and individualized communications. Even if all sensitive terms are disguised per spam, it is extremely likely that non-sensitive words in the templates are not obfuscated and are unlikely to co-occur invalid SMS. Spammers would have to devote a significant amount of time and money to tailoring each spammer (Oecd and OECD 2011).

Attempts toward Logistic Regression Algorithms have been made in the past used in the spam filtering process. Every one of these appears to work, to the best of our knowledge and was motivated by the same set of observations as ours, without explicitly stating them. However, we discovered that none of the existing methodologies are capable of sufficiently exposing the common similarities (low entropy) among campaigns as a result they all suffer (Popovac et al. 2018).

K-Nearest Neighbor

The K Nearest Neighbor algorithm falls under the Supervised Learning category. It is used for classification (most commonly) and regression. It is a versatile algorithm also used for imputing missing values and resampling datasets. As the name (K

Nearest Neighbor) suggests it considers K Nearest Neighbors to predict the new Datapoint (Kural and Demirci 2020).

These messages instruct the user to dial a specific phone number, via which an attacker obtains the user's appropriately (Pandya 2019). KNN is an algorithm that detects data theft, and generally causes havoc (Joachims 2012). It is a harmful program that infiltrates mobile devices without the user's permission. It entails sending users unsolicited links and requesting that they download the executable file, which is risky which leads to program abuse (International Conference on Cyberworlds 2005). K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and also finds intense application in pattern recognition, data mining, and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as a Gaussian distribution of the given data) (Kural and Demirci 2020; Wei and Nguyen 2020).

Algorithm:

1. The data with a linear function in the data set.
2. Set the "spam" shaped function for classifying the data.
3. Conclude the trained data based on the threshold value obtained from the function.
4. Group the new data points into an averaged network using points.
5. Repeat point till the data nodes become empty.
6. Apply the accuracy function to network groups and data sets.
7. Return the value using the min function.

Logistic Regression:

Logistic Regression Algorithmsare a popular statistical technique in Novel spam Detection using the SMS filtering process. They typically use bag-of-words features to identify spam SMS, an approach commonly used in text classification. It is a baseline technique for dealing with spam that can tailor itself to the SMS needs of individual users which give low false positive spam detection rates that are generally acceptable to users. It is one of the oldest ways of doing spam filtering (Oecd and OECD 2011; Bishara, Bishara, and University 2016). Because spectral features of classes might change over time, you won't be able to use the same class information from one image to the next. Clustering automatically splits the dataset into groups based on their similarities. Anomaly detection can discover unusual data points in the dataset. It is useful for finding fraudulent transactions. Association mining identifies sets of items that often occur together in the dataset. Latent variable models are widely used for data preprocessing. Likewise reducing the number of features in a dataset or decomposing the dataset into multiple components. (*The performance of soft computing techniques on content-based SMS spam filtering* 2015).

The spam filtering process is a sort of machine learning in which the training data is presented to the algorithm without any pre-assigned labels or scores. As a result, first is to self-discover any naturally occurring patterns in the training data set. Spam filtering process, in which the algorithm automatically groups its training examples into categories with similar features, and principal component analysis, in which the algorithm identifies which features are most useful for discriminating

between different training examples and discards the rest, are two common examples.

Algorithm:

1. Data set values are based on the total data points in the datasets.
2. Using the spam formation formula, find the count between k data points.
3. Based on the previous point, consider the point with less count.
4. Using Naive Bayes Algorithm, find total data points in each.
5. Based on maximum neighbors assign new data points and new data sets
6. Repeat point 5 for every new data point.
7. Use the accuracy function on maximum neighbors and return the value.

RESULT

In Table 1: The statistical comparison of the Effective Model Of SMS spam detection using two sample groups was done through SPSS version 21. Target and Accuracy are dependent variables and the remaining are independent variables. Analysis was done for mean, standard deviation, and independent T-test in Novel Spam Detection

In Table 2: Group Statistics T-Test Logistic Regression Algorithm with Standard Error Mean and KNN algorithm. Independent Sample T-Test is applied with the sample collections by fixing the level of significance as ($p>0.05$) with a confidence interval of 96% after applying the SPSS calculation to the KNN algorithm.

In Table 3: Independent sample tests of accuracy and Precision values are predicted and also calculate the p-value, Mean Difference, and Confidence interval. Logistic Regression Algorithms and KNN are significantly different from each other.

In fig 1: Bar chart representing the comparison of Mean Accuracy of Effective Model Of SMS Spam Detection computed with KNN and Logistic Regression Algorithm. To produce the most consistent results with minimal standard deviation. KNN algorithm appears to produce.

There is a significant difference between the LR and KNN algorithms. We will test our classification model on our prepared dataset in this final phase, as well as analyze the performance of SMS detection on our dataset. We utilize accuracy to quantify the effectiveness of classifiers in order to evaluate the performance of our constructed classification and compare it to current ways. For the SMS spam filtering process, the experiment used various classifiers such as decision trees, KNN classifiers, and unsupervised Algorithms. Among the other classifiers, the Naive Bayes Algorithm had the highest accuracy.The LR Algorithm accuracy is 96.0% and k-nearest neighbor algorithm accuracy is 89.0% and the data collection phase is used as spam.

STATISTICAL ANALYSIS

The analysis was done using IBM SPSS software. An independent sample t-test is carried out for analysis. Independent variables are datasets and the dependent variable is accuracy and different iterations done with a maximum of 35 samples each iteration was predicted the accuracy was noted for analyzing accuracy. The value obtained from the iterations of the independent sample T-test was performed. The independent values are used in the analysis of Novel Spam Detection.

DISCUSSION

In this study of Spam message detection, random selection by a LR Algorithm is

higher than k-nearest algorithms. In order to clean the dataset, data pretreatment was employed in this research for implementation. Data preprocessing involves a number of procedures, including data cleaning, data integration, data transformation, and data reduction.

A recent project aimed at determining the basic template used to create signatures. It looks at Spam filtering process feeds generated by bots and infers the basic templates from anchor texts and macros. It also reverse-engineers the header portion of the template using mail header knowledge because the input (training data) to the system is a clean trace consisting of pure spam generated from templates; this approach falls into the domain of Logistic Regression Algorithm by deriving Spam messages.

Similar work has been carried out by the author (Zhang and Wang 2009). In the Logistic Regression Algorithm, the system attempts to find the patterns directly from the example given. So, if the dataset is labeled, it is a supervised problem, and if the dataset is unlabelled, then it is an unsupervised problem. When valid parametric estimates of probability densities are unavailable or difficult to calculate (Akinyelu 2021). KNN classification was born out of the requirement to perform discriminant analysis. Fix and Hodges devised a non-parametric method for pattern classification that became known as the k-nearest neighbor rule in an unpublished US Air Force School of Aviation Medicine study in 1951. The basic principle is to assign a person to the population whose sample contains the greatest number of "K-nearest neighbors". Some of the formal features of the k-nearest neighbor were studied further in 1967. KNN classification was developed from the reliable parametric estimates of probability densities needed, discriminant analysis is required. The following are some of the various financial applications of KNN. Predict the price stock based on corporate performance metrics and economic data. Neural Network Financial risk management and understanding of Futures trading. Credit score, loan management, bank customer profiling, and money laundering investigations are just a few of the services available ("M-Spam, Spam, Spam" 2000).

CONCLUSION

Previous work on the supervised spam filtering process was discussed and the LR Algorithm process was discussed. I briefly reviewed earlier work on spam campaign detection before comparing our approach, a recently suggested anti-spam scheme that aims to deduce underlying templates from which signatures are generated. SMS spam detection system was successfully developed. The current study focused on machine learning algorithms and Logistic Regression Algorithm over KNN for higher classification in detecting messages. The Logistic Regression Algorithm accuracy is 96.0% and the k-nearest neighbor algorithm accuracy is 89.0%.

DECLARATIONS

Conflict of interests:

No conflict of interest in the manuscript.

Author contribution:

Author BNS was involved in data collection and data analysis. Author BS involved in the Action process, Data verification, Validation and Critical review of the manuscript.

Acknowledgment:

The authors would like to express their gratitude to 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 organization for providing financial support that enabled us to complete the study.

1. Staples.INC
2. SNEW.AI Technologies
3. Saveetha School of Engineering
4. Saveetha Institute of Medical and Technical Sciences
5. Saveetha University

REFERENCES

1. Akbari, Fatemeh, and Hedieh Sajedi. 2015. "SMS Spam Detection Using Selected Text Features and Boosting Classifiers." *2015 7th Conference on Information and Knowledge Technology (IKT)*. https://doi.org/10.1109/ikt.2015.7288782.
2. Akinyelu, Andronicus A. 2021. "Advances in Spam Detection for Email Spam, Web Spam, Social Network Spam, and Review Spam: ML-Based and Nature-Inspired-Based Techniques." *Journal of Computer Security*. https://doi.org/10.3233/jcs-210022.
3. Ali, Syed Sarmad, and Junaid Maqsood. 2018. ".Net Library for SMS Spam Detection Using Machine Learning: A Cross Platform Solution." *2018 15th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*. https://doi.org/10.1109/ibcast.2018.8312266.
4. Alzahrani, Amani, and Danda B. Rawat. 2019. "Comparative Study of Machine Learning Algorithms for SMS Spam Detection." *2019 SoutheastCon*. https://doi.org/10.1109/southeastcon42311.2019.9020530.
5. Bishara, Dr Saied, Saied Bishara, and University. 2016. "Special Needs Children within Regular Classes and in Separate Classes." *The International Journal of Social Sciences and Humanities Invention*. https://doi.org/10.18535/ijsshi/v3i9.08.
6. Bosaeed, Sahar, Iyad Katib, and Rashid Mehmood. 2020. "A Fog-Augmented Machine Learning Based SMS Spam Detection and Classification System." *2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC)*. https://doi.org/10.1109/fmec49853.2020.9144833.
7. Devarajan, Yuvarajan, Beemkumar Nagappan, Gautam Choubey, Suresh Vellaiyan, and Kulmani Mehar. 2021. "Renewable Pathway and Twin Fueling Approach on Ignition Analysis of a Dual-Fuelled Compression Ignition Engine." *Energy & Fuels: An American Chemical Society Journal* 35 (12): 9930–36.
8. Dhanraj, Ganapathy, and Shanmugam Rajeshkumar. 2021. "Anticariogenic Effect of Selenium Nanoparticles Synthesized Using Brassica Oleracea." *Journal of Nanomaterials* 2021 (July). https://doi.org/10.1155/2021/8115585.
9. Diale, Melvin, Turgay Celik, and Christiaan Van Der Walt. 2019. "Unsupervised Feature Learning for Spam Email Filtering." *Computers & Electrical Engineering*. https://doi.org/10.1016/j.compeleceng.2019.01.004.
10. International Conference on Cyberworlds. 2005. *2005 International Conference on Cyberworlds: CW 2005*

: [proceedings] : 23-25 November, 2005, Singapore.

11. Joachims, Thorsten. 2012. *Learning to Classify Text Using Support Vector Machines*. Springer Science & Business Media.
12. Kamath, S. Manjunath, K. Sridhar, D. Jaison, V. Gopinath, B. K. Mohamed Ibrahim, Nilkantha Gupta, A. Sundaram, P. Sivaperumal, S. Padmapriya, and S. Shantanu Patil. 2020. "Fabrication of Tri-Layered Electrospun Polycaprolactone Mats with Improved Sustained Drug Release Profile." *Scientific Reports* 10 (1): 18179.
13. Karasoy, Onur, and Serkan Ballı. 2021. "Spam SMS Detection for Turkish Language with Deep Text Analysis and Deep Learning Methods." *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-021-06187-1.
14. Kural, Oguz Emre, and Sercan Demirci. 2020. "Comparison of Term Weighting Techniques in Spam SMS Detection." *2020 28th Signal Processing and Communications Applications Conference (SIU)*. https://doi.org/10.1109/siu49456.2020.9302315.
15. Liu, Xiaoxu, Haoye Lu, and Amiya Nayak. 2021. "A Spam Transformer Model for SMS Spam Detection." *IEEE Access*. https://doi.org/10.1109/access.2021.3081479.
16. "M-Spam, Spam, Spam." 2000. *Network Security*. https://doi.org/10.1016/s1353-4858(00)12006-9.
17. Nagwani, Naresh Kumar, and Aakanksha Sharaff. 2017. "SMS Spam Filtering and Thread Identification Using Bi-Level Text Classification and Clustering Techniques." *Journal of Information Science*. https://doi.org/10.1177/0165551515616310.
18. Nandhini, Joseph T., Devaraj Ezhilarasan, and Shanmugam Rajeshkumar. 2020. "An Ecofriendly Synthesized Gold Nanoparticles Induces Cytotoxicity via Apoptosis in HepG2 Cells." *Environmental Toxicology*, August. https://doi.org/10.1002/tox.23007.
19. Oecd, and OECD. 2011. "Women Devote Most of Their Time to Physical Childcare, While Men Devote Most of Their Time to Teaching, Reading and Playing with Their Children." https://doi.org/10.1787/soc_glance-2011-graph10-en.
20. Pandya, Darshit. 2019. "Spam Detection Using Clustering-Based SVM." *Proceedings of the 2019 2nd International Conference on Machine Learning and Machine Intelligence*. https://doi.org/10.1145/3366750.3366754.
21. Parakh, Mayank K., Shriraam Ulaganambi, Nisha Ashifa, Reshma Premkumar, and Amit L. Jain. 2020. "Oral Potentially Malignant Disorders: Clinical Diagnosis and Current Screening Aids: A Narrative Review." *European Journal of Cancer Prevention: The Official Journal of the European Cancer Prevention Organisation* 29 (1): 65–72.
22. Perumal, Karthikeyan, Joseph Antony, and Subagunasekar Muthuramalingam. 2021. "Heavy Metal Pollutants and Their Spatial Distribution in Surface Sediments from Thondi Coast, Palk Bay, South India." *Environmental Sciences Europe* 33 (1).

https://doi.org/10.1186/s12302-021-00501-2.

23. Pham, Quoc Hoa, Supat Chupradit, Gunawan Widjaja, Muataz S. Alhassan, Rustem Magizov, Yasser Fakri Mustafa, Aravindhan Surendar, Amirzhan Kassenov, Zeinab Arzehgar, and Wanich Suksatan. 2021. "The Effects of Ni or Nb Additions on the Relaxation Behavior of Zr55Cu35Al10 Metallic Glass." *Materials Today Communications* 29 (December): 102909.
24. Popovac, Milivoje, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, and Andras Anderla. 2018. "Convolutional Neural Network Based SMS Spam Detection." *2018 26th Telecommunications Forum (TELFOR)*. https://doi.org/10.1109/telfor.2018.8611916.
25. Rafique, Muhammad Zubair, and Muhammad Abulaish. 2012. "Graph-Based Learning Model for Detection of SMS Spam on Smart Phones." *2012 8th International Wireless Communications and Mobile Computing Conference (IWCMC)*. https://doi.org/10.1109/iwcmc.2012.6314350.
26. Rajalingam, Mallikka. 2020. *Text Segmentation and Recognition for Enhanced Image Spam Detection: An Integrated Approach*. Springer Nature.
27. Saeedian, Mehrnoush Famil, and Hamid Beigy. 2009. "Dynamic Classifier Selection Using Clustering for Spam Detection." *2009 IEEE Symposium on Computational Intelligence and Data Mining*. https://doi.org/10.1109/cidm.2009.4938633.
28. Sasaki, M., and H. Shinnou. 2005. "Spam Detection Using Text Clustering." *2005 International Conference on Cyberworlds (CW'05)*. https://doi.org/10.1109/cw.2005.83.
29. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." *Environmental Progress & Sustainable Energy* 40 (6). https://doi.org/10.1002/ep.13696.
30. Shakiba, Tahere, Sajjad Zarifzadeh, and Vali Derhami. 2018. "Spam Query Detection Using Stream Clustering." *World Wide Web*. https://doi.org/10.1007/s11280-017-0471-z.
31. Tesfaye Jule, Leta, Krishnaraj Ramaswamy, Nagaraj Nagaprasad, Vigneshwaran Shanmugam, and Venkataraman Vignesh. 2021. "Design and Analysis of Serial Drilled Hole in Composite Material." *Materials Today: Proceedings* 45 (January): 5759–63.
32. *The performance of soft computing techniques on content-based SMS spam filtering*. 2015.
33. Uganya, G., Radhika, and N. Vijayaraj. 2021. "A Survey on Internet of Things: Applications, Recent Issues, Attacks, and Security Mechanisms." *Journal of Circuits Systems and Computers* 30 (05): 2130006.
34. Wei, Feng, and Trang Nguyen. 2020. "A Lightweight Deep Neural Model for SMS Spam Detection." *2020 International Symposium on Networks, Computers and Communications (ISNCC)*. https://doi.org/10.1109/isncc49221.2020.9297350.

35. Zhang, Hong-Yan, and Wei Wang. 2009. "Application of Bayesian Method to Spam SMS Filtering." *2009 International Conference on Information Engineering and Computer Science*. https://doi.org/10.1109/iciecs.2009.5365176.

TABLES AND FIGURES

Table 1: Comparing accuracy values with the different sample sizes. It represents the Spam activities in Novel spam detection. Data collection from N=10 sample datasets for Logistic Regression Algorithmand KNN. Data collection from N=10 sample datasets for Logistic Regression Algorithm and K-nearest Neighbors Algorithm using target variable as Independent Variable.

Sample	**Logistic Regression (Accuracy)**	**K-nearest Neighbors (Accuracy)**
1	94.0	85.0
2	93.0	93.0
3	85.0	82.0
4	88.0	84.0
5	87.0	87.0
6	88.0	88.0
7	89.0	89.0
8	90.0	90.0
9	91.0	91.0
10	94.0	92.0

Table 2: Group Statistics T-Test for Logistic Regression Algorithmwith Standard Error Mean (0.92135) and for KNN (1.05462).

Groups	**N**	**Mean**	**Std.Deviation**	**Std.Error Mean**
Logistic Regression (A)	10	90.6000	2.91357	0.92135
K-nearest Neighbor(A)	10	89.0000	3.33500	1.05462

Logistic Regression (L)	10	9.4000	2.91357	1.17568
K-nearest Neighbor(L)	10	10.0000	3.33500	1.05462

Table 3 : Independent Sample Test of Accuracy and loss of the P-value = 0.001, Significant value = 0.04, Mean Difference = 1,300 and confidence interval = (-1.6421- 4.24212). Logistic Regression Algorithm and KNN are significantly different from each other.

Accuracy and loss	**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	.308	0.04	.928	18	.366	1.30000	1.4696	-1.64212	4.24212
and not assumed (A)	-	-	.928	17.693	.366	1.30000	1.4696	-1.64212	4.24212
Equal variances assumed	.308	0.04	.928	18	.366	-1.30000	1.4696	-4.48771	1.31229
and not assumed(L	-	-	.928	17.693	.366	-1.30000	1.4696	-4.48771	1.29486

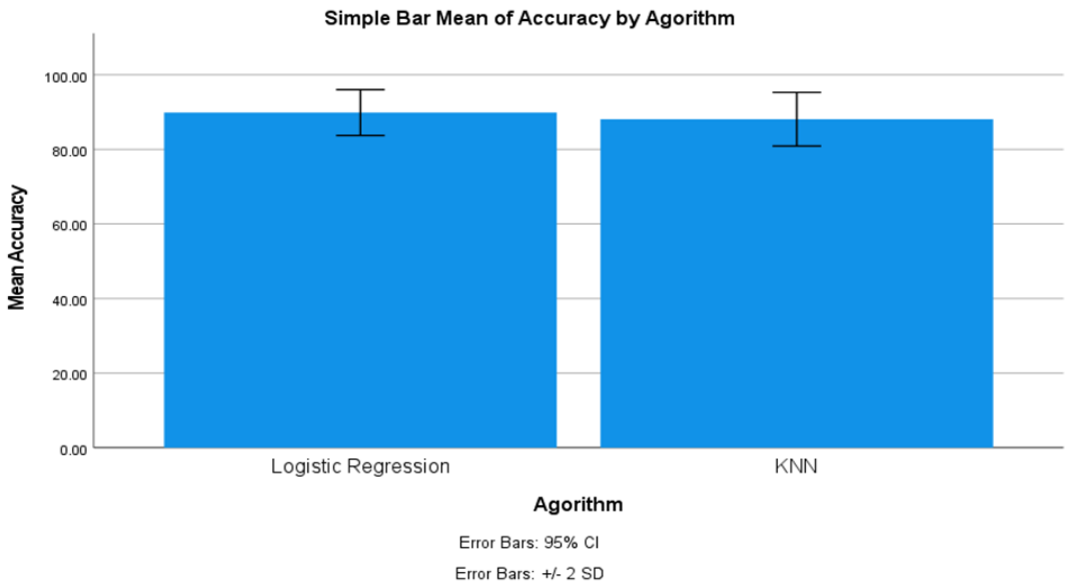

Fig 1: Comparison of Logistic Regression Algorithm and KNN Algorithms in terms of mean accuracy. It explores that the mean accuracy values and the standard deviation are moderately improved.The Logistic Regression Algorithm is slightly lower than the KNN. Graphical representation of the bar graph is plotted using X-axis Logistic Regression Algorithm vs KNN, Y-Axis displaying the error bars with a mean accuracy of detection +/- 1 SD.