Original Article

Machine Learning-based spam detection using Naïve Bayes Classifier in comparison with Logistic Regression for improving accuracy

K.Varun Kumar[1], M. Ramamoorthy[2]

[1]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
[2]Project Guide, Corresponding Author, Department of Artificial Intelligence and Machine Learning, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, TamilNadu, India, Pincode: 602105

Abstract

Aim: The aim of this research is to detect spam using machine learning with the Novel Naive Bayes Classifier (NB) and the Logistic Regression (LR). **Material and Methods:** For analyzing the spam, it needs two groups which consist of 40 samples. The two groups are group 1 which consists of Novel Naive Bayes Classifier (NB) with a sample size of 20 and group 2 which consists of Logistic Regression (LR) with a sample size of 20 and G-power (value = 0.8). **Results:** Novel Naive Bayes Classifier has an accuracy of 98.05% which is comparatively more than the Logistic Regression with an accuracy of 94.7%. The accuracy has a 2-tailed significant value of 0.012 ($p<0.05$) which is found in the Independent Sample T-Test analysis. **Conclusion:** The performance of the Novel Naive Bayes Classifier is more than the performance of Logistic Regression in terms of accuracy.

Keywords: Machine Learning, Supervised Learning, Spam detection, Ham, Novel Naive Bayes Classifier, Logistic Regression.

DOI: 10.47750/pnr.2022.13.S04.061

INTRODUCTION

Nowadays, email enables users to communicate faster with other users or with organizations and vice versa. In organizations, all the work details were shared through email only. So Spam mails or gratuitous emails cause a distraction to the users in the communication. The spam mails or gratuitous emails may contain an advertisement or unwanted information which may also include malicious code or malicious software which will help the hackers or the cyberpunks to sneak your information and make their business with our information (Trivedi 2016). A survey has reported that on average a person will receive 25-40 mails every day and 60-70% of business emails are spam mails only. Email spam is growing rapidly which causes inconvenience to users in the form of distraction, occupies more storage, wastes time and energy, increasing network traffic (Agarwal and Kumar 2018). There are a lot of tools and techniques to detect spam and to filter based on ham (consists of legitimate words) and spam (consists of gratuitous words) (Gibson et al. 2020). Spam detection is used in Google Mails, Microsoft Outlook, Yahoo mails, Internet Service Providers (ISPs), and Small-Medium sized Business organizations (SMBs) to safeguard their employee's data and networks (Cichosz 2015).

Over the past 5 years, on spam detection using machine learning, 18,400 articles have been published in Google Scholar, 27 journal papers are available in IEEE Xplore, 1,773 articles are available in ScienceDirect. There are many machine learning methods used by researchers to get rid of spam. Some methods used are: using Logistic Regression, which is a supervised learning technique. It is used to predict discrete values such as True or False, 0 or 1. Logistic regression is used to measure the probability of occurrence of an event by squeezing it into the logistic function. It shows '0' for ham and '1' for the spam (Dedeturk and Akay 2020). Another method is by using Random Forest, which is a supervised learning technique. It consists of many individual decision trees, and each tree consists of votes which are given based on the overall classification of the set of data. This algorithm chooses the individual tree with the most votes. It is the superset of the Decision Tree (Devi 2018; Wang et al. 2015). Another method is by using KNN, which is a supervised learning technique. The K-Nearest Neighbor algorithm is the process of making clusters based on similar properties. The clusters are taken as classes and it defines the class as the yield or not yield. From the K classes of documents, it has to calculate the score. The

scores are taken into consideration only if they are greater than the threshold value following the statistical value of the class (Ren and Shi 2016; Laksono, Basuki, and Bachtiar 2020). Another method is by using the Decision Tree algorithm, which is a supervised learning technique. It works well for both categorical and continuous variables. It consists of the root node as an input variable. The recursive partitioning of subsets was done until the subset values matched the target variables (Chakraborty and Mondal 2012; Saini 2021).

Our team has extensive knowledge and research experience that has translate into high quality publications (Bhansali et al. 2021; Jayanth et al. 2021; Sudhakar, Ravel, and Perumal 2021; Sathiyamoorthi et al. 2021; Deepanraj et al. 2021; Raju et al. 2021; Arun Prakash et al. 2020; Kamath et al. 2020; Shanmugam et al. 2021; Rajasekaran et al. 2020; Adhinarayanan et al. 2020; Rajesh et al. 2020; Aurtherson et al. 2021). The lacuna in the existing system is accuracy. Due to less accuracy in the existing system, some spam emails or gratuitous emails are coming into the inbox of an email which causes distraction to the user. So it wants to improve the accuracy of the proposed system using machine learning. The aim of this research is to develop machine learning-based spam detection using the Novel Naive Bayes Classifier in comparison with the Logistic Regression to improve accuracy. The goal of this research is to increase accuracy in detecting spam.

Materials And Methods

The research work is accomplished in the Image Processing Lab, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences where the laboratory provides a high infrastructure system to obtain the experimental results. There are two groups required for this research and 20 samples are required for each group (Wei 2018). Group 1 consists of Novel Naive Bayes Classifier and group 2 consists of Logistic Regression. The calculation is performed utilizing G-power 80% (Gibson et al. 2020) with alpha worth 0.05 and beta worth is 0.95 with a confidence interval of 95%.

The dataset named spam is downloaded from the Kaggle website. The spam dataset is in the form of a CSV file with a size of 491 KB which consists of ham and spam words. The ham and spam words help us to decide whether the email is a ham(legitimate mail) or spam(gratuitous mail).

Google Colab is used as the implementation for this work. All my codes are executed in Google Colab only. Google Colab is used to execute all python codes with zero configuration and free access to GPUs.It is used to combine all the executable python codes and it is divided into cells. The hardware configuration used in this research was a system with 8 GB RAM and 1.8 TB ROM of a 64-bit operating system of Windows 10 with a processor of Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.70 GHz.

Novel Naive Bayes Classifier

The Novel Naive Bayes Classifier is a supervised learning algorithm. It uses the Bayes theorem to calculate the probability of an event. It follows the properties like strong independence, easily handles large datasets, and depends on probability distribution (Cichosz 2015; Pooja and Bhatia 2018)**.** The Bayes theorem is used to calculate the probability distribution from the frequency of the dataset. From the probability distribution, the class having the highest posterior probability is chosen by the NB classifier. The posterior probability equation is as shown in equation (1) (Agarwal and Kumar 2018).

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)} \quad (1)$$

Where

A=(a_1,a_2,a_3,.....), B=(b_1,b_2,b_3,.....)

$P(A)$is evidence probability,

$P(B)$is prior probability,

$P(A|B)$ is a conditional probability,

$P(B|A)$ is posterior probability.

Pseudocode

Input: Spam dataset

Output: Accuracy of spam detection

Step-1: Initially, all of us have to download and install all the packages and libraries.

Step-2: Import all packages that are downloaded.

Step-3: It needs to load the dataset and extract the ham and spam keywords.

Step-4: Clean the dataset which includes removing single letter words, truncating white spaces, tokenizing each and every message, deleting all punctuations, changing all the letters to lowercase, etc.

Step-5: Then split the dataset into test and train datasets.

Train: X_train, Y_train.

Test: X_test, Y_test
Step-6: Train the machine which is spam and ham when they triggered the spam and ham words.
Step-7: Load the Novel Naive Bayes classifier and train the model with the training dataset.
NB=NaiveBayesClassifier()
NB.fit(X_train,Y_train)
Step-8: Calculate the probability distribution $P(B|A)$ for every class using the Bayes theorem.
Step-9: Calculate the confusion matrix and find the Accuracy.
accuracy = sum(X_test.Label ==Y_ test.predicted)/len(X_test)

Logistic Regression

Logistic Regression is a supervised learning algorithm. It is used to calculate discrete values such as 0 or 1 and True or False. It is used to find the probability of occurrence of an event by squeezing it into the logistic function. It shows 0 for ham and 1 for spam (Gupta et al. 2018). The logistic function is an S-shaped curve which is a sigmoid curve, it is shown in equation (2).

$$f(x) = \frac{M}{(1+e^{-t(x-x_0)})} \forall x \in (-\infty, \infty) \quad (2)$$

Where
M is the curve maximum value,
x, x_0 are the midpoints of the sigmoid curve,
t is the growth rate of the curve.

Pseudocode
Input: Spam dataset
Output: Accuracy of spam detection

Step-1: Download and install all the libraries required for this model.
Step-2: Import all the libraries.
Step-3: Load the dataset and extract the ham and spam keywords.
Step-4: Now clean the dataset, it includes removing all the white spaces, tokenizing each
and every message, removing all punctuations, lower case every one of the letters, truncating all the single letter words('a', 'i'), etc.
Step-5: Encode the ham as '0' and spam as '1'.
Step-6: Now divide the dataset into a test and train dataset.
Train: P, Q.
Test: P_Test, Q_Test.
Step-7: Load the Logistic Regression model and train the model with a training dataset.
LR= LogisticRegression()
LR.fit(P.iloc[:,6:],Q)
Step-8: Calculate the confusion matrix for the training dataset.
confusion_matrix(Q,LR.predict(P.iloc[:,6:]))
Step-9: Finally, calculate the confusion matrix for the test dataset.
confusion_matrix(Q_Test,LR.predict(P_Test.iloc[:,6:]))
Step-10: Predict the accuracy from the confusion matrix of the test dataset.
LR.score(P_Test.iloc[:,6:], Q_Test)

Statistical Analysis

The IBM SPSS Version 28 software is used to calculate the statistical variables like mean, standard deviation, standard error mean, mean difference, sig, and F value. The Independent Sample T-Test analysis is carried out in this research. The spam dataset with 4862 ham words and 728 spam words is used to detect spam. The dependent variables are spam and ham. The independent variables are accuracy and count of words (Udge et al. 2019).

Results

The accuracy is taken as a measurement for comparing the Novel Naive Bayes Classifier and Logistic Regression. The spam dataset is required for both comparing and analyzing the models. From the results, it is clearly shown that the novel Naive Bayes Classifier has more accuracy than Logistic Regression. The Mean accuracy for the Novel Naive Bayes Classifier is 98.05% and the Mean Accuracy for Logistic Regression is 94.7%.

Table 1 represents the statistical variables like Mean, Standard Deviation, and Standard Error Mean are measured for both the Novel Naive Bayes Classifier and the Logistic Regression. The Mean Accuracy of the Novel Naive Bayes Classifier is 98.05% which is greater than the Mean accuracy of Logistic Regression with 94.7%. The Standard Deviation of the Novel Naive Bayes Classifier(0.759) is slightly less than the Standard Deviation of Logistic Regression(0.801).
Table 2 represents the Independent Sample T-Test which consists of Mean Difference and Standard Error Difference calculated for both the Novel Naive Bayes Classifier and the Logistic Regression with Confidence Interval of 95%. The 2-tailed significant value of accuracy for both models is 0.012 ($p<0.05$).
Fig 1 depicts the Bar graph for comparing the Mean Accuracy of the Novel Naive Bayes Classifier and Logistic Regression along with error bars of Confidence intervals of 95% and SD. The X-axis consists of groups(NB, LR) and the Y-axis consists of Mean Accuracy. The Mean Accuracy of the Novel Naive Bayes Classifier is 98.05% which is slightly greater than the Mean Accuracy of Logistic Regression with 94.7%.

Discussion

The data is evaluated in IBM SPSS software with version 28 and an Independent Sample T-Test is carried out. From the results, it is clearly observed that the Novel Naive Bayes Classifier has more accuracy than the Logistic Regression. The Mean Accuracy of the Novel Naive Bayes Classifier is 98.05% which is slightly greater than the Mean Accuracy of Logistic Regression with 94.7%.

Similar findings related to the Novel Naive Bayes Classifier are (Cichosz 2015; Pooja and Bhatia 2018). In those articles, the spam dataset is selected. The dataset is cleaned and split into train and test datasets. Loading the model and training it with a training dataset. Then test the model with a test dataset and calculate the confusion matrix. The Naive Bayes Classifier offers certain advantages, such as the capacity to forecast class data fast and easily, the ability to cope with independent assumptions, and the ability to operate with categorical rather than numerical input elements (Cichosz 2015). The disadvantages are as follows: The Naive Bayes Classifier, which assumes all predictors are independent, faces the 'zero-frequency problem,' which allocates likelihood '0' to categorical variables that are available in the test dataset but not in the training dataset (Trivedi 2016). Opposite findings for spam detection are (Singh, Pamula, and Shekhar 2018). This paper makes use of the spam dataset. The data was cleaned and subdivided into two categories: training and test datasets. In this example, they used the Support Vector Machine method, which is a sort of supervised learning technology. It shows the data as a point in n-dimensional space from the input. Each feature is a value for a certain coordinate. After that, the categorization is completed by creating a hyperplane with vectors that clearly identify the two classes. This model is loaded, trained, and tested using train and test datasets. The calculations for accuracy and the confusion matrix are done.

The limitations in spam detection are making a good text classification, link analysis will be done accurately, avoiding picking ham mail to the trash instead of spam, requiring more keywords, and providing data security to users. In the future, all of us will try to increase the detection of spam to greater than 99%, and also all of us will try to decrease the error rate to less than 1%. All of us will also try to implement this spam detection by using other Deep Learning algorithms.

Conclusion

This review shows that the accuracy of the Novel Naive Bayes classifier is more when compared to the Logistic Regression model in detecting spam. The accuracy of the Novel Naive Bayes Classifier is 98.05% which is comparatively more than the accuracy of the Logistic Regression model with 98.70%.

DECLARATION

Conflict of Interest

No conflict of interest in this manuscript.

Authors Contribution

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

Acknowledgments

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

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

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

References

1. Adhinarayanan, Rajesh, Aravindh Ramakrishnan, Gopal Kaliyaperumal, Melvinvíctor De Poures, Rajesh Kumar Babu, and Damodharan Dillikannan. 2020. "Comparative Analysis on the Effect of 1-Decanol and Di-N-Butyl Ether as Additive with diesel/LDPE Blends in Compression Ignition Engine." *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, June, 1–18.
2. Agarwal, Kriti, and Tarun Kumar. 2018. "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization." *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. https://doi.org/10.1109/iccons.2018.8662957.
3. Arun Prakash, V. R., J. Francis Xavier, G. Ramesh, T. Maridurai, K. Siva Kumar, and R. Blessing Sam Raj. 2020. "Mechanical, Thermal and Fatigue Behaviour of Surface-Treated Novel Caryota Urens Fibre–reinforced Epoxy Composite." *Biomass Conversion and Biorefinery*, August. https://doi.org/10.1007/s13399-020-00938-0.
4. Aurtherson, P. Babu, Bhanu Teja Nalla, Karthikeyan Srinivasan, Kulmani Mehar, and Yuvarajan Devarajan. 2021. "Biofuel Production from Novel Prunus Domestica Kernel Oil: Process Optimization Technique." *Biomass Conversion and Biorefinery*, May. https://doi.org/10.1007/s13399-021-01551-5.
5. Bhansali, Karan J., Kamlesh R. Balinge, Subodh U. Raut, Shubham A. Deshmukh, M. Senthil Kumar, C. Ramesh Kumar, and Pundlik R. Bhagat. 2021. "Visible Light Assisted Sulfonic Acid-Functionalized Porphyrin Comprising Benzimidazolium Moiety for Photocatalytic Transesterification of Castor Oil." *Fuel* 304 (November): 121490.
6. Chakraborty, Sarit, and Bikromadittya Mondal. 2012. "Spam Mail Filtering Technique Using Different Decision Tree Classifiers through Data Mining Approach - A Comparative Performance Analysis." *International Journal of Computer Applications in Technology* 47 (16): 26–31.
7. Cichosz, Paweł. 2015. "Naïve Bayes Classifier." *Data Mining Algorithms: Explained Using R*, January, 118–33.
8. Dedeturk, Bilge Kagan, and Bahriye Akay. 2020. "Spam Filtering Using a Logistic Regression Model Trained by an Artificial Bee Colony Algorithm." *Applied Soft Computing*. https://doi.org/10.1016/j.asoc.2020.106229.
9. Deepanraj, B., N. Senthilkumar, D. Mala, and A. Sathiamourthy. 2021. "Cashew Nut Shell Liquid as Alternate Fuel for CI Engine—optimization Approach for Performance Improvement." *Biomass Conversion and Biorefinery*, February. https://doi.org/10.1007/s13399-021-01312-4.
10. Devi, Khongbantabam Susila. 2018. "Random Forests Spam Email Classification System." *Journal of Computer Engineering & Information Technology*. https://doi.org/10.4172/2324-9307.1000190.
11. Gibson, Simran, Biju Issac, Li Zhang, and Seibu Mary Jacob. 2020. "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms." *IEEE Access*. https://doi.org/10.1109/access.2020.3030751.
12. Gupta, Mehul, Aditya Bakliwal, Shubhangi Agarwal, and Pulkit Mehndiratta. 2018. "A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers." *2018 Eleventh International Conference on Contemporary Computing (IC3)*. https://doi.org/10.1109/ic3.2018.8530469.
13. Jayanth, Bellappu Venkat, Melvin Victor Depoures, Gopal Kaliyaperumal, Damodharan Dillikannan, Dilipsingh Jawahar, Kumaran Palani, and Ganesha Prasad Meravanigee Shivappa. 2021. "A Comprehensive Study on the Effects of Multiple Injection Strategies and Exhaust Gas Recirculation on Diesel Engine Characteristics That Utilize Waste High Density Polyethylene Oil." *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, June, 1–18.
14. Kamath, Manjunath, Subha Krishna Rao, Jaison, Sridhar, Kasthuri, Gopinath, Sivaperumal, and Shantanu Patil. 2020. "Melatonin Delivery from PCL Scaffold Enhances Glycosaminoglycans Deposition in Human Chondrocytes – Bioactive Scaffold Model for Cartilage Regeneration." *Process Biochemistry* 99 (December): 36–47.
15. 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.
16. . Pooja, and Komal Kumar Bhatia. 2018. "Spam Detection Using Naive Bayes Classifier." *International Journal of Computer Sciences and Engineering*. https://doi.org/10.26438/ijcse/v6i7.712716.
17. Rajasekaran, S., D. Damodharan, K. Gopal, B. Rajesh Kumar, and Melvin Victor De Poures. 2020. "Collective Influence of 1-Decanol Addition, Injection Pressure and EGR on Diesel Engine Characteristics Fueled with diesel/LDPE Oil Blends." *Fuel* 277 (October): 118166.
18. Rajesh, A., K. Gopal, De Poures Melvin Victor, B. Rajesh Kumar, A. P. Sathiyagnanam, and D. Damodharan. 2020. "Effect of Anisole Addition to Waste Cooking Oil Methyl Ester on Combustion, Emission and Performance Characteristics of a DI Diesel Engine without Any Modifications." *Fuel* 278 (October): 118315.
19. Raju, P., K. Raja, K. Lingadurai, T. Maridurai, and S. C. Prasanna. 2021. "Glass/Caryota Urens Hybridized Fibre-Reinforced nanoclay/SiC Toughened Epoxy Hybrid Composite: Mechanical, Drop Load Impact, Hydrophobicity and Fatigue Behaviour." *Biomass Conversion and Biorefinery*, March. https://doi.org/10.1007/s13399-021-01427-8.
20. Ren, Biyi, and Yuliang Shi. 2016. "Research On Spam Filter Based On Improved Naive Bayes and KNN Algorithm." *Proceedings of the 2016 4th International Conference on Machinery, Materials and Computing Technology*. https://doi.org/10.2991/icmmct-16.2016.220.
21. Saini, Anshul. 2021. "Decision Tree Algorithm - A Complete Guide - Analytics Vidhya." August 29, 2021. https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/.
22. 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.
23. Shanmugam, Rajasekaran, Damodharan Dillikannan, Gopal Kaliyaperumal, Melvin Victor De Poures, and Rajesh Kumar Babu. 2021. "A Comprehensive Study on the Effects of 1-Decanol, Compression Ratio and Exhaust Gas Recirculation on Diesel Engine Characteristics Powered with Low Density Polyethylene Oil." *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects* 43 (23): 3064–81.
24. Singh, Manmohan, Rajendra Pamula, and Shudhanshu Kumar Shekhar. 2018. "Email Spam Classification by Support Vector Machine." *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*.

https://doi.org/10.1109/gucon.2018.8674973.
25. Sudhakar, M. P., Merlyn Ravel, and K. Perumal. 2021. "Pretreatment and Process Optimization of Bioethanol Production from Spent Biomass of Ganoderma Lucidum Using Saccharomyces Cerevisiae." *Fuel* 306 (December): 121680.
26. Trivedi, Shrawan Kumar. 2016. "A Study of Machine Learning Classifiers for Spam Detection." *2016 4th International Symposium on Computational and Business Intelligence (ISCBI)*. https://doi.org/10.1109/iscbi.2016.7743279.
27. Udge, Ganesh, Mahesh Mohite, Shubhankar Bendre, Yogeshwar Birnagal, and Disha Wankhede Mrs. 2019. "Statistical Analysis for Twitter Spam Detection." *International Journal of Scientific Research in Science, Engineering and Technology*, May, 624–29.
28. Wang, W. B., F. Yin, H. Sun, and P. Li. 2015. "Random Forest Algorithm for Spam Filtering Based on Machine Learning." In *Electronic Engineering and Information Science*, 225–28. CRC Press.
29. Wei, Qijia. 2018. "Understanding of the Naive Bayes Classifier in Spam Filtering." https://doi.org/10.1063/1.5038979.

ABLES AND FIGURES

Table 1. The statistical calculations for the Novel Naive Bayes Classifier and Logistic Regression are measured. The mean accuracy of the Novel Naive Bayes classifier is 98.05 and the Mean accuracy of Logistic Regression is 94.70. Standard Deviation for Novel Naive Bayes Classifier is 0.759 and Standard Deviation for Logistic Regression is 0.801.

	Groups	**N**	**MEAN**	**Std. Deviation**	**Std. Error Mean**
Accuracy	NB	20	98.05	0.75915	0.16975
	LR	20	94.70	0.80131	0.17918

Table 2. Statistical Independent T-Test between Novel Naive Bayes Classifier and Logistic Regression with a confidence interval of 95%. The 2-tailed significant value of accuracy is 0.012 ($p<0.05$).

		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**
Accuracy	Equal variances assumed	.937	.339	13.573	38	.012	3.350	.2468	2.8503	3.8496
	Equal variances not assumed			13.573	37.89	.012	3.350	.2468	2.8502	3.8497

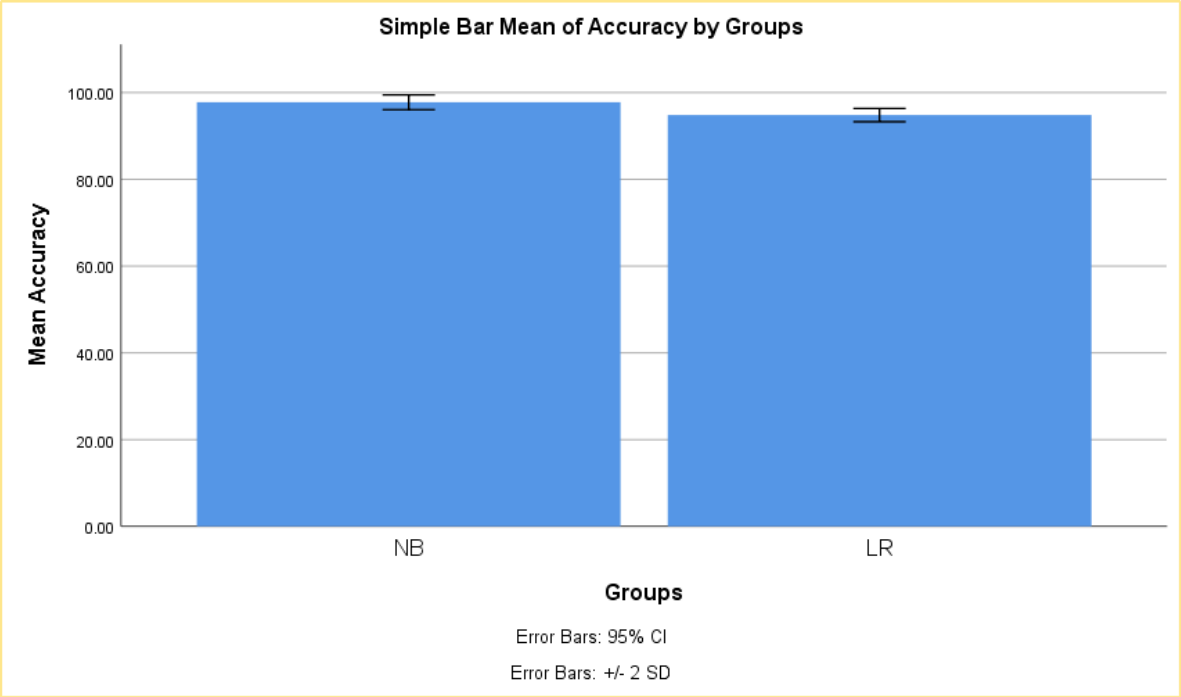

Fig. 1. Bar Graph for comparing the Novel Naive Bayes Classifier and the Logistic Regression with 95% confidence interval and with SD of +/- 2. The mean accuracy of the Novel Naive Bayes Classifier is slightly greater than the Logistic Regression. The X-axis consists of groups (NB, LR) and the Y-axis consists of Mean Accuracy.