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): 349-356
DOI: 10.2478/bjlp-2022-004037

SMS Spam Detection Using Multinational Naive Bayes Algorithm Compared with Decision Tree Algorithm

J. Krishna Prasad
Research Scholar, Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nādu, India. Pin code: 602105.

S. Christy
Project Guide, Corresponding Author, Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nādu, India. Pin code: 602105.

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

Abstract

Aim: The main objective of this research is to improve accuracy through machine learning algorithms. Multinational Naive Bayes Algorithm and Decision Tree Algorithm were used in this research. **Materials and Methods:** Detection is performed by Multinational Naive Bayes Algorithm (N=10) over Decision Tree Algorithm (N=10). Sample size is calculated using GPower with pretest power as 0.8 and alpha 0.05. **Result:** Mean performance of Multinational Naive Bayes Algorithm (97.80%) is high compared to Decision Tree Algorithm (96.50%). Significance value for performance and loss is 0.398 (P>0.536). **Conclusion**: The mean performance of a Novel SMS spam detection using Multinational Naive Bayes Algorithm is better than Decision Tree Algorithm.

Keywords

Multinational Naive Bayes Algorithm, Decision Tree Algorithm, SMS, Novel Spam detection, Accuracy.

INTRODUCTION

The SMS spam problem is gradually increasing in text messaging. Many users do not want spam messages in their mobile phones (Yang et al. 2006). There are many popular text classification techniques to solve SMS spam problems (Salehi 2011). Text classification techniques include various techniques like Logistic Regression, Naive Bayes Classifier, Nearest Neighbors, Support Vector Machine (Narayan, Rout, and Jena 2018). Decision Tree, Neural Networks and Rule Induction. By using these techniques filtration of SMS is done based on text classification techniques (Abdulhamid et al. 2017). In today's world, along with the drop in SMS tariffs, there is a growth in SMS spam, which is exploited by some persons as a substitute for advertising and fraud. As a result, it becomes a critical issue because it can annoy and harm users, and one solution is with automatic SMS spam filtering (Suleiman and Al-Naymat 2017). In this research, the main application is that the ideal minimal support and comparison study of the use of Naive Bayes alone with the use of Naive Bayes and FP-Growth collaboration are carried out (Bosaeed, Katib, and Mehmood 2020). Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique

- The applications of SMS spam detection are to improve SMS spam filtering performance by integrating two data mining tasks: association and classification for web applications, information systems, to detect spam and ham messages more accurately. A novel approach that uses machine learning classification algorithms to detect and filter spam messages (Choudhary and Jain 2017).

SMS spam detection has been carried out by researchers and 16 research articles are published in IEEE Digital Explore and 70 research articles are published in Google Scholar in the year 2017-2021. SMS Spam Filtering Using Supervised Machine Learning Algorithms (17 Citations). This article determines that SVM algorithm gives highest accuracy followed by Naive Bayes Algorithm for filtering the messages (Navaney, Dubey, and Rana 2018). Enhancing Spam Detection on Mobile Phone Short Message Service (SMS) Performance using FP-Growth and Naive Bayes Classifier (26 citations). This article determines that the adoption of minimum support helps in reducing the difficulties associated with dealing with limited features owing to the limited number of characters in SMS (Arifin, Shaufiah, and Bijaksana 2016). Spam detection using Multinational Naive Bayes Algorithm and decision tree mechanism in social networks (10 citations). This article uses the Weka tool. Performance metrics such as TP Rate, FP Rate, Precision, Recall, F-Measure, and Class are used to assess the effectiveness of the proposed mechanism. Multinational Naive Bayes Algorithm and Decision Tree algorithms are used to detect spam and ham messages which results in Multinational Naive Bayes Algorithm performing better than the Decision Tree algorithm (Goyal, Chauhan, and Parveen 2016). spam Detection Approach for Secure Mobile Message Communication Using Machine Learning Algorithms (7 citations). This article determines that it achieved high performance in terms of accuracy (Shah Nazir et al 2020). Overall, the article's best study is Enhancing Spam Detection on Mobile Phone Short Message Service (SMS) Performance using FP-Growth and Naive Bayes Classifier (Arifin, Shaufiah, and Bijaksana 2016).

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 drawback of the existing system of SMS spam detection has a lower predictions and accuracy rate and takes more time. Though much research has been carried out in this field, the accuracy level is low. Accordingly, a technique with more accuracy level is to be determined. The aim of this research work is to predict the best technique to improve accuracy and predictions of SMS Spam Detection Using Multinational Naive Bayes Algorithm Compared with Decision Tree Algorithm.

MATERIALS AND METHODS

The study setting was done in the Data Analytics Lab, Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The dataset used in this study is taken from kaggle (https://github.com/mohitgupta-omg/Kaggle-SMS-Spam-Collection-Dataset-/blob/master/spam.csv). The total number of groups for this project are 2. The total number of samples is 304. The sample size for each group is 152. G power 0.8. In SMS spam detection, to modify the problem of low accuracy rate Multinational Naive Bayes Algorithm and Decision Tree Algorithm were used (Trần 2018).

Multinational Naive Bayes Algorithm

Multinational Naive Bayes Algorithm is used in this work as sample preparation for group 1 has a sample size of 152. The pseudocode for Multinational Naive Bayes Algorithm is shown in Table 1. Multinational naive bayes algorithm is a simple machine learning algorithm that can be applicable in a variety of circumstances, most prominently in spam classification. Multinational naive bayes is a classification method based on Bayes Theorem

and the assumption of predictor independence. multinational Naive Bayes is mostly utilized in text classification with a large training dataset. multinational Naive Bayes aids in the development of rapid machine learning models capable of making quick predictions. Mean accuracy of Naive Bayes Algorithm is 97.80%. The multinational naive bayes algorithm is calculated by using the formula,

$$P(A|B) = P(B|A) * P(A) / P(B) \quad (1)$$

Decision Tree Algorithm

Decision Tree Algorithm is used in this work as sample preparation for group 2 has a sample size of 152. The pseudocode for the decision tree algorithm is shown in Table 2. Decision Tree Algorithm is a Supervised learning approach that may be used to solve classification and regression issues, however it is most commonly used to solve classification problems. Decision tree is one of the predictive modeling techniques used in statistics, data mining, and machine learning is the decision tree. Decision trees are built using an algorithmic technique that discovers alternative ways to segment a data set depending on certain parameters. Decision tree is one of the most extensively used and useful supervised learning algorithms. Decision Trees are a non-parametric supervised learning approach that may be used for classification as well as regression applications. It is used to extract behavioral patterns of spam.Mean accuracy of Decision Tree Algorithm is 96.5%.

In order to simulate a multinational naive bayes algorithm, first open the Jupiter tool, upload the dataset, check the null values, count all the messages, find the count and unique messages, using vectorizer prices fit the data in it, use the Naive Bayes Algorithm code, run the code and finally get the output accuracy values. Similarly in order to simulate Decision Tree Algorithm, first open the Jupyter tool, upload the dataset, check the null values, count all the messages, find the count and unique messages, using vectorizer prices fit the data in it, use the Decision Tree Algorithm code, run the code and finally get the output accuracy values. The decision tree algorithm is calculated by using the formula,

$$Gini = 1 - \Sigma_{i=1}^{c} (Pi)2 \quad (2)$$

Statistical Analysis

The analysis was finished by IBM SPSS adaptation 21. In SPSS, datasets are prepared using 10 as sample size for both the multinational naive bayes algorithm and Decision Tree Algorithm. Group id is given as 1 for multinational naive bayes algorithm and 2 for Decision Tree Algorithm, group id is given as a grouping variable and accuracy is given as a testing variable. The attributes are spam messages, ham messages, time. Dependent variables are accuracy and loss. Independent variables are actual messages, received time, date, type of messages. Independent t test is carried out in this research work.

RESULTS

The overall sample size employed in statistical tools is 152. This data is analyzed using the multinational naive bayes algorithm and the Decision Tree Algorithm. Both the multinational naive bayes algorithm and the Decision Tree Algorithm are subjected to statistical data analysis. System group and accuracy values are being computed. These 152 data samples for each method, together with their losses, are also utilized to produce statistical values for comparison. The group statistics table displays the number of samples gathered. The mean and standard deviation are computed and entered, as well as the accuracies.

Table 1, shows the steps involved in the multinational naive bayes algorithm. Table 2, shows the steps involved in the decision tree algorithm. Table 3, shows the Accuracy and loss for SMS Spam Detection using multinational naive bayes algorithm. Table 4, shows

the Accuracy and loss for SMS spam detection using Decision Tree Algorithm. Table 5, shows group statistics values along with mean, standard deviation and standard error mean for the two algorithms are also specified. Independent sample T test is applied for data set fixing confidence interval as 95%. Table 6, shows independent t sample tests for algorithms. The comparative accuracy analysis, mean of loss between two algorithms are specified. Figure 1 shows comparison of mean of accuracy and mean loss between multinational naive bayes algorithm and decision tree algorithm.

DISCUSSION

The accuracy of Decision Tree Algorithm is 96.50% whereas multinational naive bayes algorithm has higher accuracy of 97.80% with p = 0.536 which shows that multinational naive bayes algorithm is better than Decision Tree Algorithm. Mean, standard deviation and standard mean values for multinational naive bayes algorithm are 89.8860, 5.95987, 1.88468 respectively. Table 5 specifies accuracy and loss of multinational naive bayes algorithm. Similarly, for Decision Tree Algorithm mean, standard deviation and standard mean values are 86.3230, 6.77101, 2.14118 respectively. Table 6 specifies accuracy and loss Decision Tree Algorithm.

This research increases prediction for accuracy to find better ways to SMS spam detection in accordance with their data (Popovac et al. 2018). This model has a slow processing rate with better accuracy (Alzahrani and Rawat 2019). Slow processing rate is due to usage of a large database but in case of a smaller database, both the processing and accuracy are faster and better. Above problem's complexity will be reduced once a model is built (Liu, Lu, and Nayak 2021). Despite various fact that many researchers have discovered various recognized models, many of them are unable to accurately perform better algorithms (Shahi and Shakya 2018). Many applications can be developed to predict accuracy from various platforms (Akbari and Sajedi 2015).

CONCLUSION

From this study of SMS spam detection, the mean accuracy of decision tree algorithm is 96.50% whereas multinational naive bayes algorithm has a higher mean accuracy of 97.80%. Hence it is inferred that multinational naive bayes algorithm appears to be better in accuracy when compared to Decision Tree Algorithm.

DECLARATIONS

Conflict of Interest

No conflict of interest in this manuscript.

Authors Contribution

Author KP was involved in data collection, data analysis and manuscript writing. Author SC was involved in conceptualization, data validation and critical reviews 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 necessary infrastructure to carry out this work successfully.

Funding

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

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

REFERENCES

Abdulhamid, Shafi'i Muhammad, Muhammad Shafie Abd Latiff, Haruna Chiroma, Oluwafemi Osho, Gaddafi Abdul-Salaam, Adamu I. Abubakar, and Tutut Herawan. 2017. "A Review on Mobile SMS Spam Filtering Techniques." *IEEE Access*. https://doi.org/10.1109/access.2017.2666785.

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.

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.

Arifin, Dea Delvia, Shaufiah, and Moch Arif Bijaksana. 2016. "Enhancing Spam Detection on Mobile Phone Short Message Service (SMS) Performance Using FP-Growth and Naive Bayes Classifier." *2016 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*. https://doi.org/10.1109/apwimob.2016.7811442.

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.

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.

Choudhary, Neelam, and Ankit Kumar Jain. 2017. "Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique." *Communications in Computer and Information Science*. https://doi.org/10.1007/978-981-10-5780-9_2.

Goyal, Saumya, R. K. Chauhan, and Shabnam Parveen. 2016. "Spam Detection Using KNN and Decision Tree Mechanism in Social Network." *2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. https://doi.org/10.1109/pdgc.2016.7913250.

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.

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.

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.

Narayan, Rohit, Jitendra Kumar Rout, and Sanjay Kumar Jena. 2018. "Review Spam Detection Using Semi-Supervised Technique." *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-981-10-3376-6_31.

Navaney, Pavas, Gaurav Dubey, and Ajay Rana. 2018. "SMS Spam Filtering Using Supervised Machine Learning Algorithms." *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. https://doi.org/10.1109/confluence.2018.8442564.

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.

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.

Salehi, Saber. 2011. *A Comparative Evaluation of Machine Learning Approaches in SMS*

Spam Detection.
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.
Shahi, Tej Bahadur, and Subarna Shakya. 2018. "Nepali SMS Filtering Using Decision Trees, Neural Network and Support Vector Machine." *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. https://doi.org/10.1109/icacccn.2018.8748286.
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.
Suleiman, Dima, and Ghazi Al-Naymat. 2017. "SMS Spam Detection Using H2O Framework." *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2017.08.335.
Trần, Hữu Trung. 2018. *SMS Spam Detection for Vietnamese Messages: Graduation Thesis for the Honor Degree of Information Technology*.
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.
Yang, Zhen, Xiangfei Nie, Weiran Xu, and Jun Guo. 2006. "An Approach to Spam Detection by Naive Bayes Ensemble Based on Decision Induction." *Sixth International Conference on Intelligent Systems Design and Applications*. https://doi.org/10.1109/isda.2006.253725.

TABLES AND FIGURES

Table 1. Steps involved in SMS spam detection for Multinational Naive Bayes Algorithm

INPUT: Training datasets for SMS Spam detection
Step 1: Importing Required Libraries Step 2: Reading Dataset Step 3: Data Preprocessing Step 4: Tokenizing Step 5: Model building Step 6: Predicting validation of data
OUTPUT: Spam messages are detected and obtained accuracy

Table 2. Steps involved in SMS spam detection for Decision Tree Algorithm

INPUT: Training datasets for SMS spam detection
Step 1: Importing Required Libraries Step 2: Reading Dataset Step 3: Data Preprocessing Step 4: Tokenizing Step 5: Model building Step 6: Predicting validation of data
OUTPUT: spam messages are detected and obtained accuracy

Table 3. Accuracy and loss for SMS Spam Detection using Multinational Naive Bayes Algorithm

Iteration	Accuracy	Loss
1	80.00	20.00
2	82.00	18.00
3	86.00	14.00
4	87.50	12.50
5	89.65	10.35
6	91.72	8.28
7	93.87	6.13
8	94.32	5.68
9	96.00	4.00
10	97.80	2.20

Table 4. Accuracy and loss for SMS Spam Detection using Decision Tree Algorithm

Iteration	Accuracy	Loss
1	76.80	23.20
2	78.54	21.46
3	80.63	19.37
4	82.41	17.59
5	84.41	15.37
6	87.25	12.75
7	90.52	9.48
8	92.64	7.36
9	93.31	6.69
10	96.50	3.50

Table 5. Group Statistical analysis for Multinational Naive Bayes and Decision Tree Algorithm: Mean, Standard Deviation and standard error, mean are determined

	Group	N	Mean	Std Deviation	Std. Error Mean
Accuracy	Multinational Naive Bayes	10	89.8860	5.95987	01.88468

	Decision Tree	10	86.3230	6.77101	2.14118
Loss	Multinational Naive Bayes	10	10.1140	5.95987	1.88468
	Decision Tree	10	13.6770	6.77101	2.14118

Table 6. Independent sample T-test t is performed on two groups for significance and standard error determination. P value is greater than 0.05 (0.536) and it is considered to be statistically insignificant with 95% confidence interval

		Levene's test for Equality of variance		**T-Test for equality of mean**						
				t	**df**	**Sig(2-tailed)**	**Mean difference**	**Std.Error Difference**	**95% confidence of Difference**	
		F	**Sig**						**Lower**	**Upper**
Accuracy	**Equal variances assumed**	0.398	0.536	1.249	18	0.228	3.56300	2.85248	-2.42984	9.55584
	Equal Variances not assumed	-	-	1.249	17.715	0.228	3.56300	2.85248	-2.43677	9.56277
Loss	**Equal variances assumed**	0.398	0.536	-1.249	18	0.228	-3.56300	2.85248	-9.55584	2.42984
	Equal Variances not assumed	-	-	-1.249	17.715	0.228	-3.56300	2.85248	-9.56277	2.43677

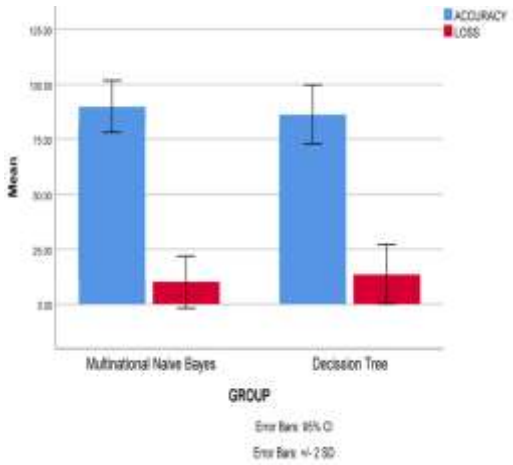

Fig. 1. Comparison of Naive Bayes Algorithm and Decision Tree Algorithm in terms of accuracy. The mean accuracy of Naive Bayes Algorithm is greater than Decision Tree Algorithm and the standard deviation is also slightly higher than Decision Tree Algorithm. X-axis: Naive Bayes Algorithm vs Decision Tree Algorithm. Y-axis: Mean accuracy of detection $\pm$ 2 SD.