Science and Technology Publishing (SCI & TECH)
ISSN: 2632-1017
Vol. 7 Issue 2, February - 2023

Machine Learning-Based Spam Filter Model For Sms Messages

Gloria Ngozi Ezeh[1]
Information Technology Department, School of Information and Communication Technology,
Federal University of Technology Owerri.
gloriaezeh2014@yahoo.com

Michael Paul Esu[2]
Department Of Electrical/Electronic and Computer Engineering
University of Uyo, Akwa Ibom State Nigeria

Ofonime Dominic Okon[3]
Department Of Electrical/Electronic and Computer Engineering
University of Uyo, Akwa Ibom State Nigeria

***Abstract*— In this paper, spam filter model for SMS using machine learning is presented. The dataset used in the study consists of a large text file that contains 5574 SMS messages where the data on each of the messages combines the content of the SMS messages with a label that designates the message as either a spam or ham (that is, legitimate). After pre-processing the dataset, the following machine learning algorithms were used to train the model, namely; Naive Bayes, Support Vector Machine (SVM), and Logistic Regression algorithm. The model was trained using a CPU from Google Colab. The dataset was split into the training set and test set. About 20% of the training data were used for validation set. After training and validation, the model was evaluated on the test set. The metrics used for the assessment of the models are accuracy, precision, recall, and f1-score. Also, confusion matrix was used to measure the performance of the machine learning classification algorithms. In all, by considering the two Natural Language Processing (NLP) techniques used , namely; Bag of Words and Term Frequency Inverse Document Frequency (TF-IDF), the results showed that the machine learning models trained using the Bag of Words model performed better than those trained using TF-IDF model. The best performing learning-Based model is the SVM Model with the linear kernel. It has an accuracy score of 97.30% and an f1-score of 0.8929.**

Keywords:* *Naive Bayes, Natural Language Processing, Support Vector Machine, Bag of Words, Logistic Regression algorithm, Google Colab, Confusion Matrix

1. Introduction

Across the globe, Short Message Service (SMS) has become widely adopted and it has become one of the dominant communication channels among mobile device users [1,2,3,4,5,6,7,8,9,10]. According to the International Telecommunication Union (ITU) publication, an average of 2000 SMS/day and over 14 trillion SMS/ year was estimated for the 2013 [11]. The relative low cost of SMS services has been attributed to its wide application. Notably, SMS does not require internet connection. In addition, the ease of use and ease of access of SMS message on mobile device has prompted many people to prefer SMS over email [12,13,14,15,16,17].

Due to the teeming population of SMS users, SMS-based service attacks rate of has also grown tremendously [18,19,20,21,22,23]. Among the various attacks on mobile devices, SMS spam attack is the most dominant [24,25,26,27]. Specifically, spam SMS message is unsolicited and in most cases fake or unwanted text messages delivered to a mobile phone. In most cases, the spam SMS messages are indiscriminately sent using bulk SMS mechanism without the authorization of the recipients [28,29,30,31].

Furthermore, the mobile phone users are increasing engaging in services that require SMS as a communication channel. Notable services in this category includes, Facebook Messenger [32,33,34], mobile and regular banking applications alert/notification system [35,36,37,38], iMessage [39,40,41], e-government platforms, among others. In response, spam attack has also taken advantage of such relevant services to launch spam attacks on the subscribers. This has also given rise to increase in difficulty of managing SMS spam attacks as subscribers finds it difficult in some cases to determine the genuine SMS messages. With the increasing incidence of SMS spam attacks, and the growing difficulty in determining the genuine SMS messages, many subscribers are most likely to fall victim to these spam messages.

Accordingly, this paper is aimed at addressing the problem by developing a machine learning-based model for detecting spam SMS messages [42,43,44,45,46]. The machine learning mechanism can then be interfaced with a mobile responsive progressive web app [47,48,49] which will enable the mobile device user to detect and delete or isolate the spam SMS messages in real-time. In this paper, a case study dataset of SMS messages and three different machine learning algorithms are used to train the machine learning-based spam filter model, namely; Naive Bayes, Support Vector Machine (SVM) [50,51,52,53], and Logistic Regression algorithm [54,55,56,57]. Also, two Natural Language Processing (NLP) techniques used, namely; Bag of Words (BoW) [58,59,60,61] and Term Frequency Inverse Document Frequency (TF-IDF) were also employed in the model development [62,63,64,65]. After training and validation, the model was evaluated on test dataset. The two main metrics used for the assessment of the models are accuracy and precision. Furthermore, other metrics like confusion matrix, as well as recall and f1-score were also used.

2. Methodology

2.1 Dataset

The study utilized a dataset which has a large text file that contains several lines of text where each of those lines corresponds to a text message. Notably, the dataset contains 5574 SMS messages, where the data on each of the messages combines the content of the SMS messages with a label that designates the message as either a spam or ham (that is, legitimate).

The structure of the dataset is such that there are five columns (Figure 1), where the column denoted as v1 is the label that shows whether the SMS message is a "ham" or "spam". The column denoted as v2 is the content of the SMS message. The other three columns denoted as "Unnamed" are not needed in building the model, hence, they are dropped. In order to improve understandability of the dataset, the code snippet in Figure 2 is employed to drop the unwanted columns and also to rename the retained columns; the snapshot of the resulting dataset is given in Figure 3.

Apart from the dataset, other requisite libraries imported and used in this project includes; numpy, pandas and matplotlib.pyplot. The Pie chart representation of the distribution of the dataset into Ham and Spam in data components is given in Figure 4. Particularly, the case study dataset consist of about 13.0% of spam SMS messages and the 86.6% of ham messages. In addition, the ham message length fails in the range of around 30-40 characters while the spam message length is in the range of 155-160 characters.

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN
...	...	...	...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...	NaN	NaN	NaN
5568	ham	Will Ì_ b going to esplanade fr home?	NaN	NaN	NaN
5569	ham	Pity, * was in mood for that. So...any other s...	NaN	NaN	NaN
5570	ham	The guy did some bitching but I acted like i'd...	NaN	NaN	NaN
5571	ham	Rofl. Its true to its name	NaN	NaN	NaN

5572 rows × 5 columns

Figure 1: Snapshot of the Dataset

```
# remove last three columns
dataset = dataset.drop (labels=['Unnamed: 2',
'Unnamed: 3', 'Unnamed: 4'], axis=1)
Dataset
# Rename columns
dataset.columns = ['label', 'text']
dataset
```

Figure 2 The code snippet used to improve understandability of the dataset

	label	text
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will ì_ b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So...any other s...
5570	ham	The guy did some bitching but I acted like i'd...
5571	ham	Rofl. Its true to its name

Figure 3: Snapshot of the Dataset with updated and labelled columns

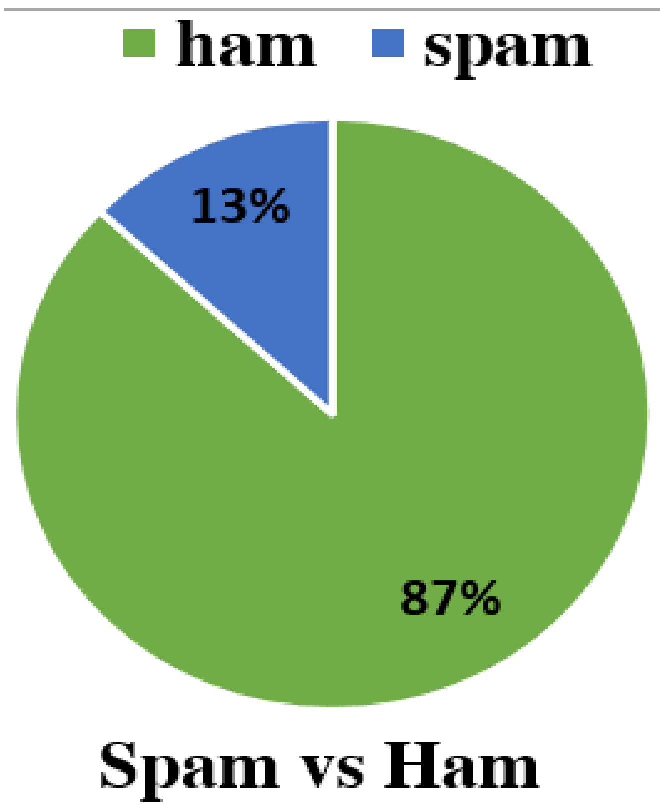

Figure 4: Pie chart representation of the distribution of the dataset into Ham and Spam in data components

2.2 Data Pre-processing

Pre-processing of the data was carried out on the dataset. First, in the dataset each of the messages is divided alphanumeric characters. At this point special characters are eliminated from the message text feature space. Also removed from the message text are dots and space. When there is non-alphanumeric characters existing within a set of contiguous characters the entire alphabetic string is saved in the memory as token.

The message content was processed using Regular Expressions (Regex). The Regex operation was used to make the emails and web addresses, as well as the phone numbers and other numbers in the text file to be in a uniform encode symbols and also to remove punctuation and white spaces and eventually convert all the characters to lower case. Also, word stemming was performed which is essentially extraction of the base form of the words. The BoW model was created for use in the extraction of features from the text. At this point, the BoW model was trained with each of the following machine learning algorithms: (i) Naive Bayes ML Classifier Algorithm (ii) Logistic Regression ML Algorithm (iii) Support Vector Machine

ML Algorithm. In this paper, the F1 score performance metric is used to evaluate the performance of the models. Based on its mode of operation, the F1 score is at its best with a value of 1 and at its worst with a value of 0. Again, in this paper, the metric considers the precision along with recall to compute the score.

Apart from Bag of words model, another Natural Language Processing (NLP) used is the Term Frequency Inverse Document Frequency (TF-IDF) which is employed to assess the importance of the various words in the text. It simply identifies how relevant a word is. The output from the TF-IDF operation was also trained with each of the following machine learning algorithms: (i) Naive Bayes ML Classifier Algorithm (ii) Logistic Regression ML Algorithm (iii) Support Vector Machine ML Algorithm. Again, model performance was evaluted

2.4 Model training and performance evaluation

After pre-processing the dataset, different machine learning algorithms were used to train the model. The model was trained using a CPU from Google Colab. The machine learning algorithms used were Naive Bayes, Support Vector Machine (SVM), and Logistic Regression. The dataset was split into the training set and test set. About 20% of the training data were used for validation set. After training and validation, the model was evaluated on the test set. The two main metrics used for the assessment of the models are accuracy and precision. Furthermore, other metrics like confusion matrix, as well as recall and f1-score were also used. The f1-score is a trade-off between precision and recall.

The confusion matrix gives a count of the number of true positives, false positives, true negatives, and false negatives. The true positives are the messages the ML model correctly classified as spam. The false positives are the messages that were wrongly classified as spam. The true negatives are the messages that were correctly classified as ham, while the false negatives are the messages wrongly classified as ham.

3. Results and Discussion

3.1 Results of obtained for the Bag of Words Model

The results on the results on the various performance metrics obtained for the bag of words model are shown in the Table 1. The accuracy plot for the ML algorithms using BoW Model is shown in Figure 5, the precision plot is shown in Figure 6, the recall plot is shown in Figure 7 and the F1-Score plot for the ML algorithms using Bag of Words Model is shown in Figure 8.

Table 1. The results on Accuracy, Precision, Recall and F1-Score obtained for the BoW model

ML Algorithm	Accuracy	Precision	Recall	F1-Score
Naive Bayes	87.53%	0.5196	0.8859	0.6550
Logistic Regression	97.93%	1.0000	0.8456	0.9163
Support Vector Machine (Linear Kernel)	98.39%	1.0000	0.8792	0.9357
Support Vector Machine (RBF Kernel)	97.85%	1.0000	0.8389	0.9124
Support Vector Machine (Poly Kernel)	93.72%	0.9877	0.5369	0.6956

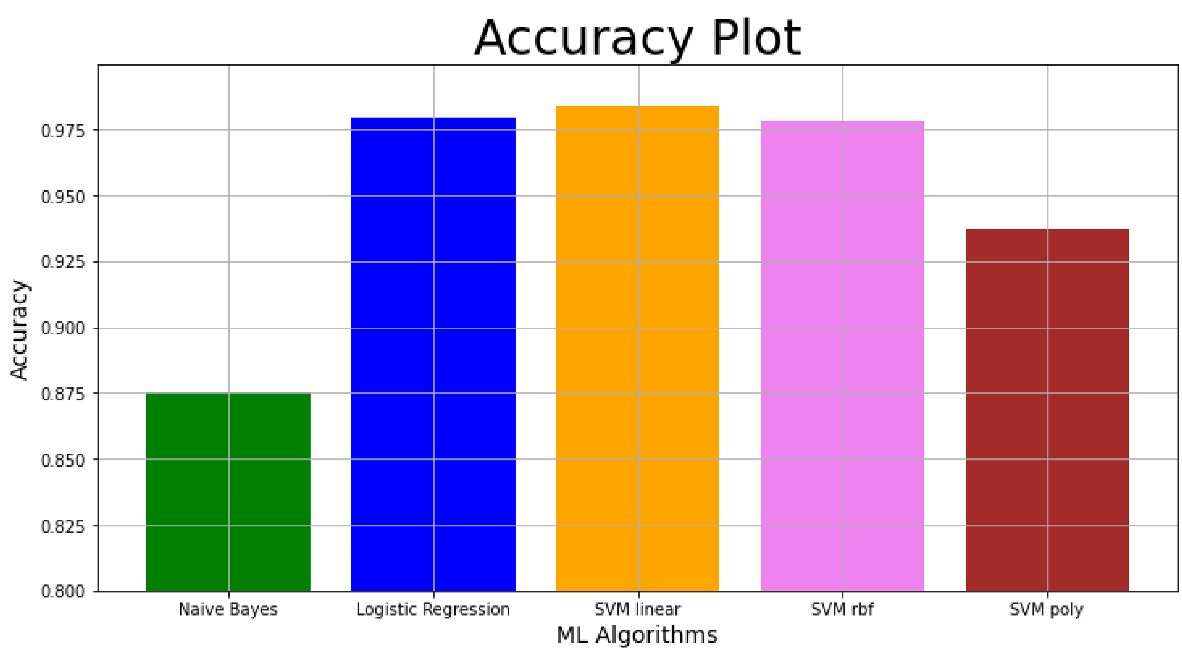

Figure 5: Accuracy Plot for ML Algorithms using BoW Model

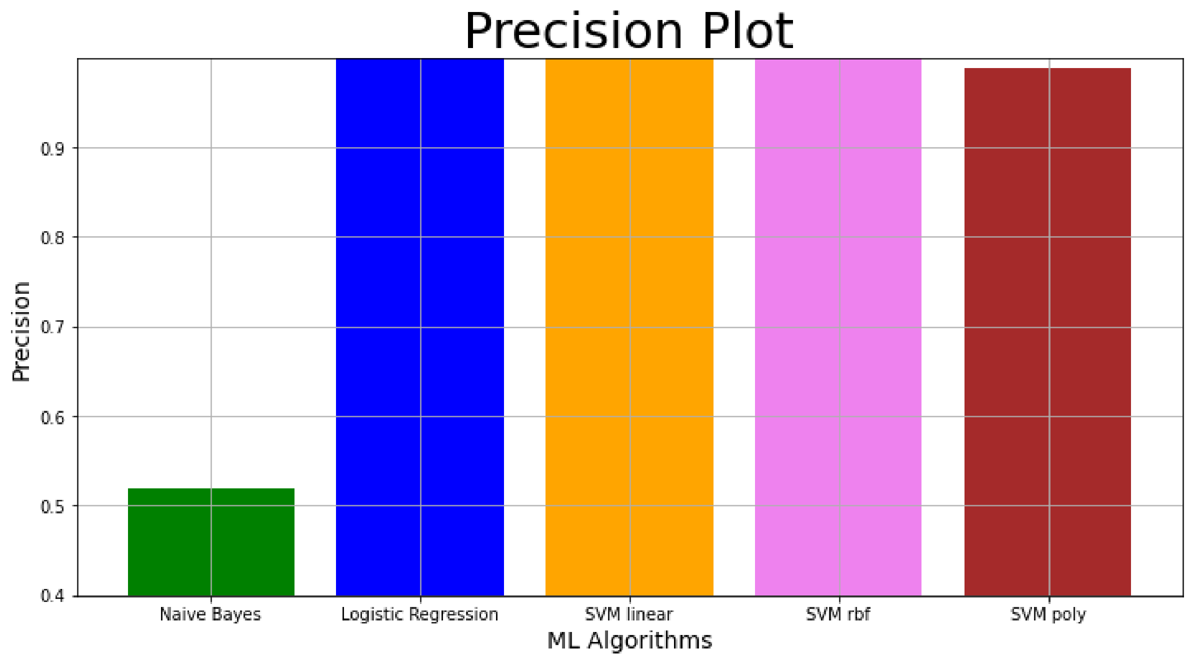

Figure 6: Precision Plot for ML Algorithms using BoW Model

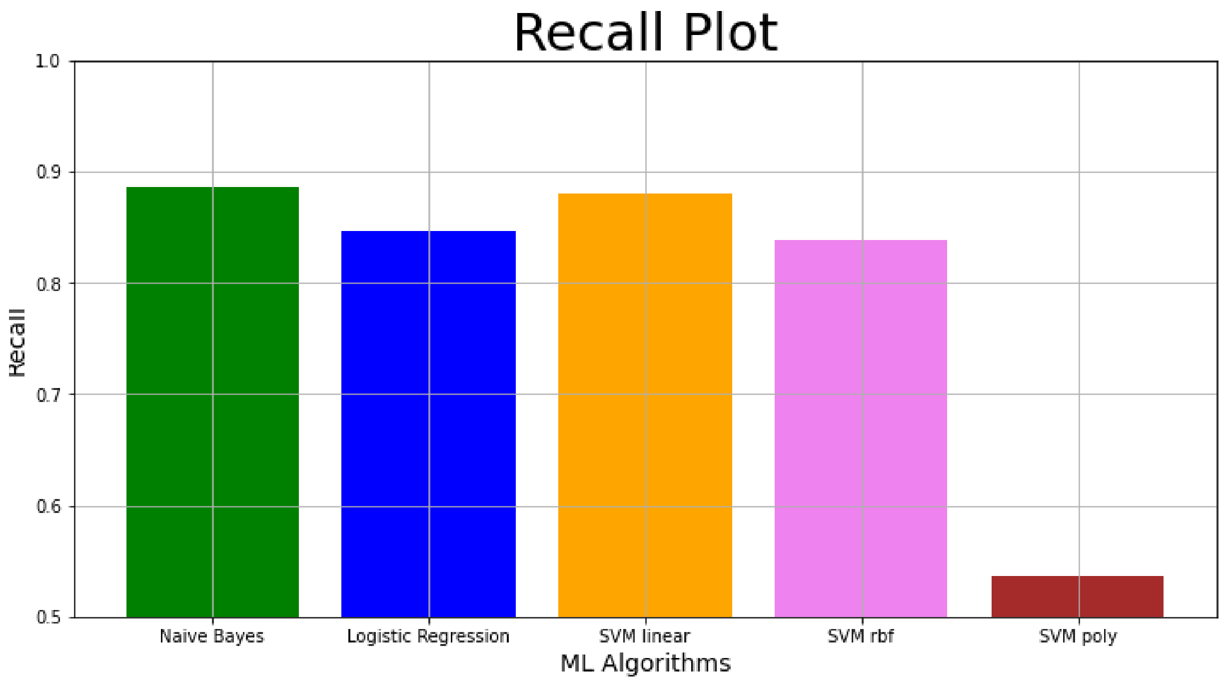

Figure 7: Recall Plot for ML Algorithms using BoW Model

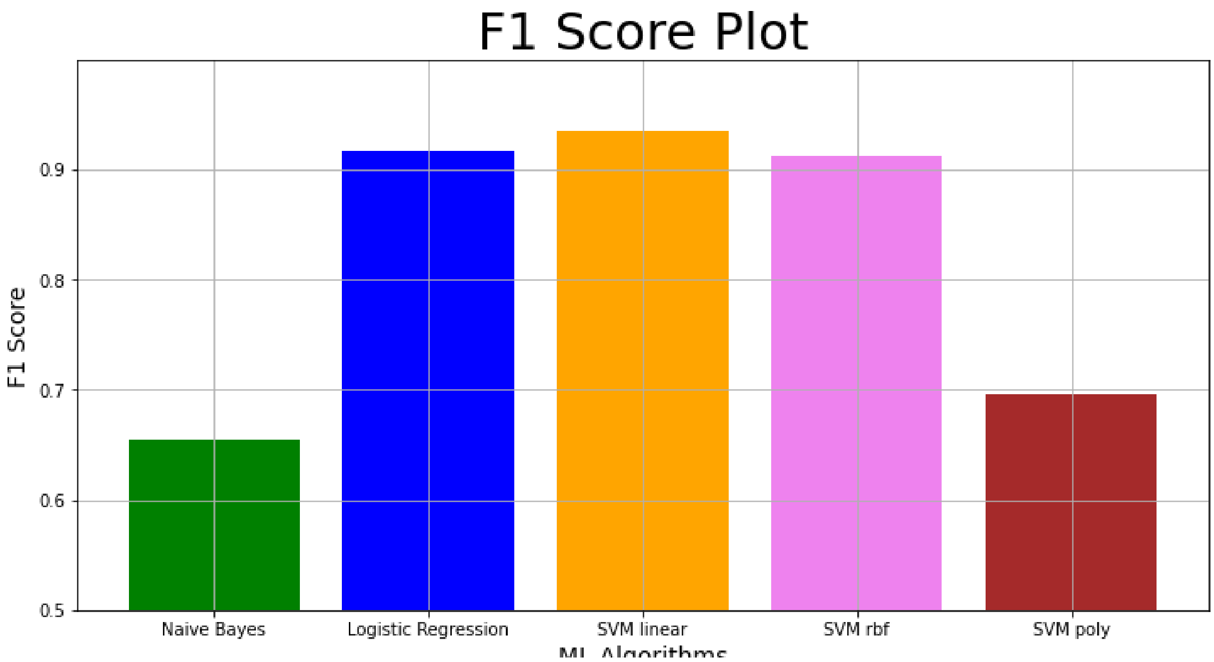

Figure 8: F1-Score Plot for ML Algorithms using BoW Model

The result of the confusion matrix for each machine learning algorithm is shown in the Table 2. The True Positives for the ML Algorithms using BoW Model is shown in Figure 9, the False is shown in Figure 10, the True Negatives is shown in Figure 11 and the False Negatives for ML the algorithms using BoW Model is shown in Figure 12.

Table 2. True Positives, False Positives, True Negatives, and False Negatives result for BoW Model

ML Algorithm	True Positives	False Positives	True Negatives	False Negatives
Naive Bayes	844	17	132	122
Logistic Regression	966	23	126	0
Support Vector Machine (Linear Kernel)	966	18	131	0
Support Vector Machine (RBF Kernel)	966	24	125	0
Support Vector Machine (Poly Kernel)	965	69	80	1

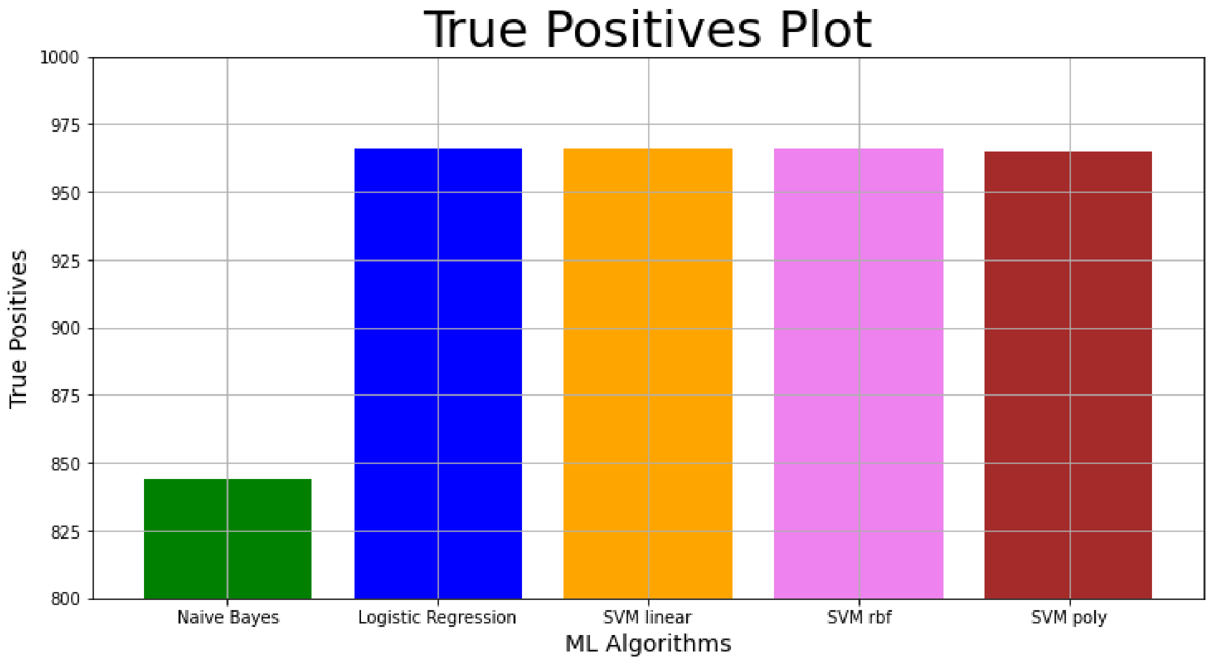

Figure 9: True Positives for ML Algorithms using BoW Model

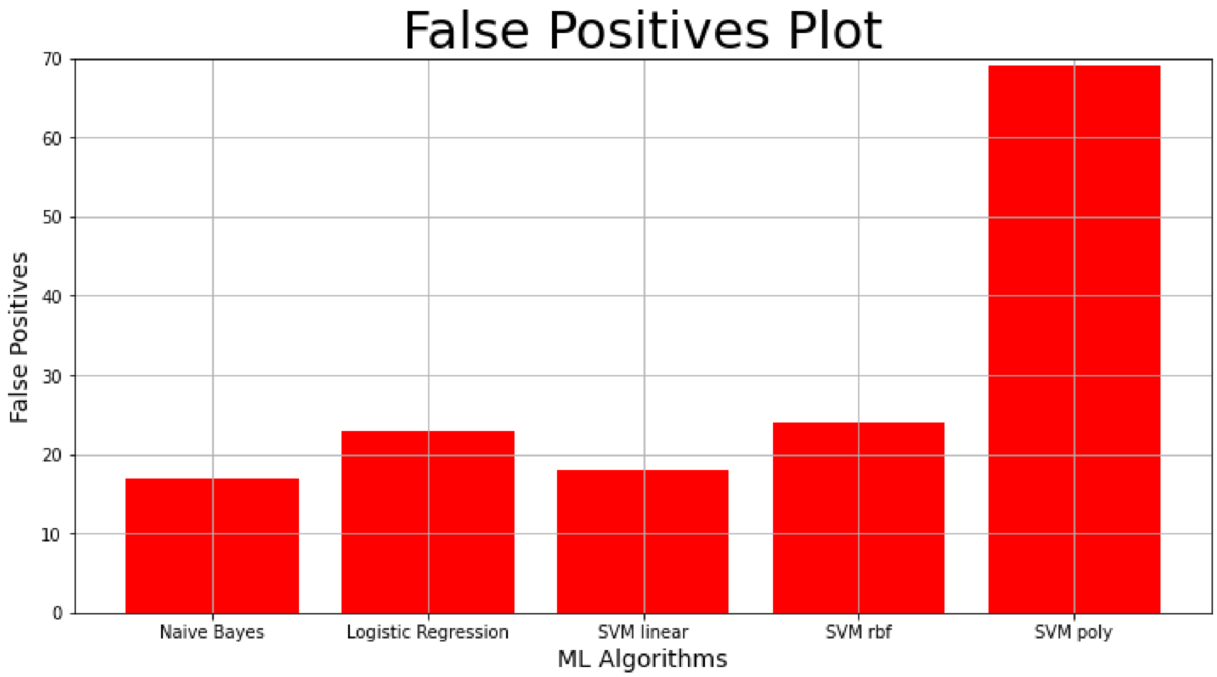

Figure 10: False Positives for ML Algorithms using BoW Model

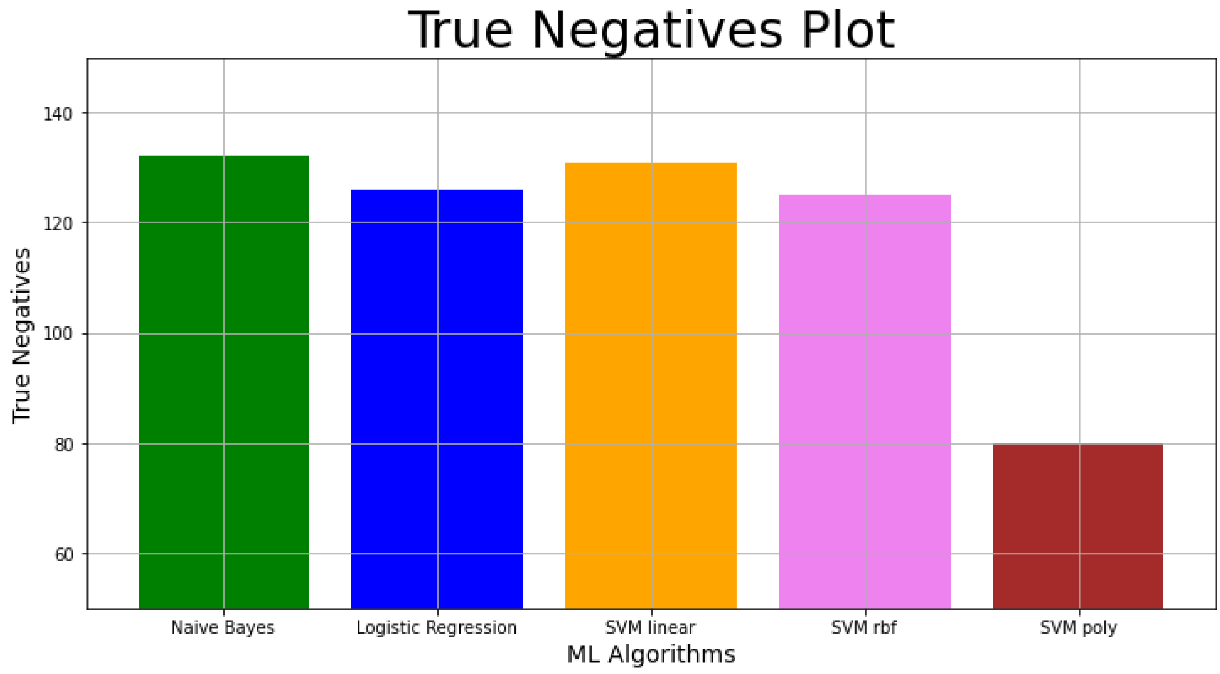

Figure 11: True Negatives for ML Algorithms using BoW Model

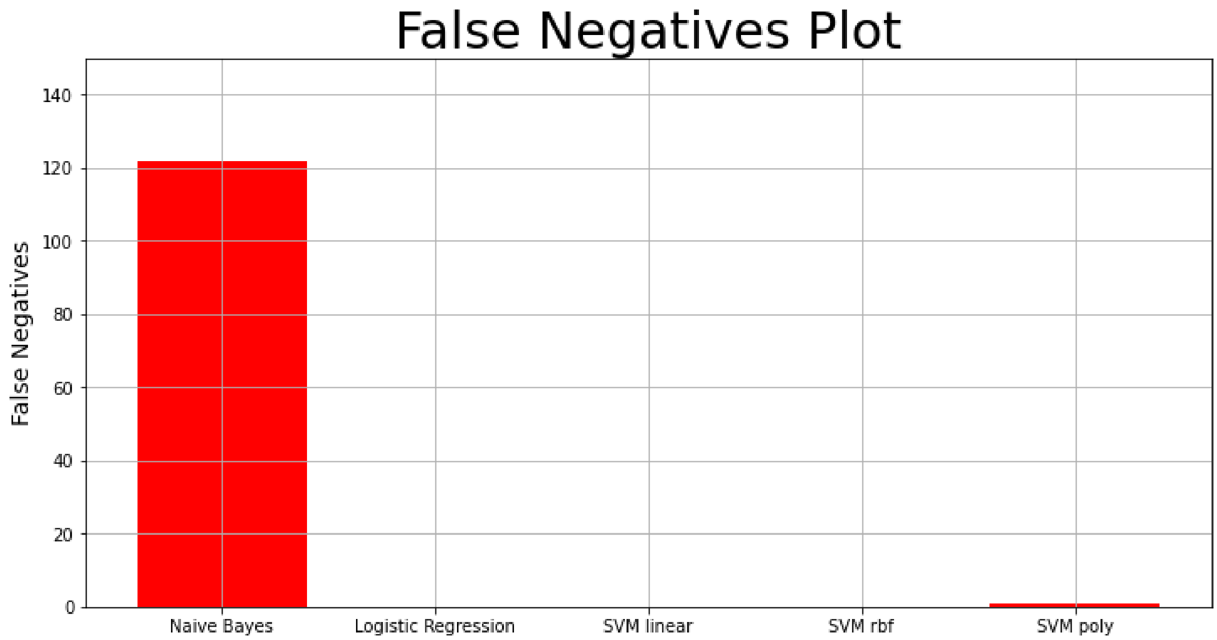

Figure 12: False Negatives for ML Algorithms using BoW Model

3.2 Results of the Term Frequency Inverse Document Frequency (TF-IDF) Model

The results on the various performance metrics obtained for the TF-IDF model are shown in the Table 3. The accuracy plot for the ML algorithms using TF-IDF Model is shown in Figure 13, the precision plot is shown in Figure 14, the recall plot is shown in Figure 15 and the F1-Score plot for the ML algorithms using TF-IDF Model is shown in Figure 16.

The result of the confusion matrix for each machine learning algorithm is shown in the Table 4 for the TF-IDF Model. The True Positives for the ML Algorithms using TF-IDF Model is shown in Figure 17, the False is shown in Figure 18, the True Negatives is shown in Figure 19 and the False Negatives for ML the algorithms using TF-IDF Model is shown in Figure 20.

In all, by considering the two Natural Language Processing (NLP) techniques used , namely; BoW and TF-IDF, the results showed that the machine learning models trained using the boW model performed better than those trained using TF-IDF model. The best performing learning-Based model is the SVM Model with the linear kernel. It has an accuracy score of 97.30% and an f1-score of 0.8929.

Table 3: The results on the various performance metrics obtained for TF-IDF Model

ML Algorithm	Accuracy	Precision	Recall	F1-Score
Naive Bayes	87.00%	0.5080	0.8523	0.6366
Logistic Regression	95.34%	0.9709	0.6711	0.7937
Support Vector Machine (Linear Kernel)	97.30%	0.9542	0.8389	0.8929
Support Vector Machine (RBF Kernel)	96.68%	0.9828	0.7651	0.8604
Support Vector Machine (Poly Kernel)	92.38%	1.0000	0.4295	0.6009

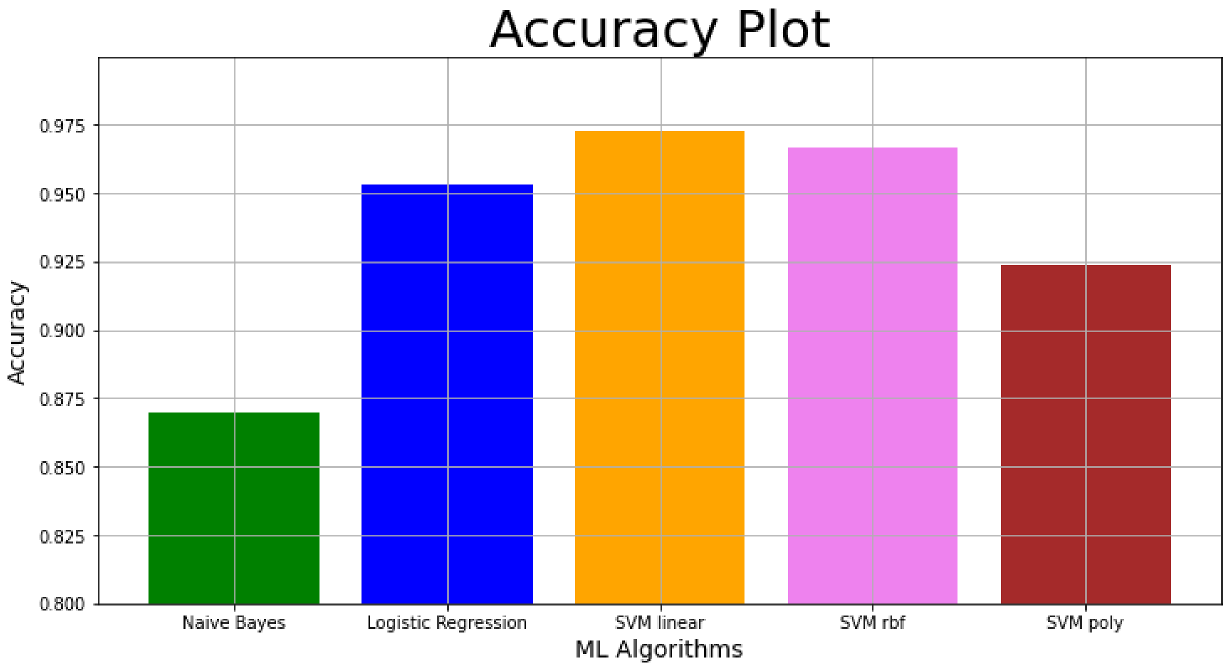

Figure 13: Accuracy Plot for ML Algorithms using TF-IDF Model

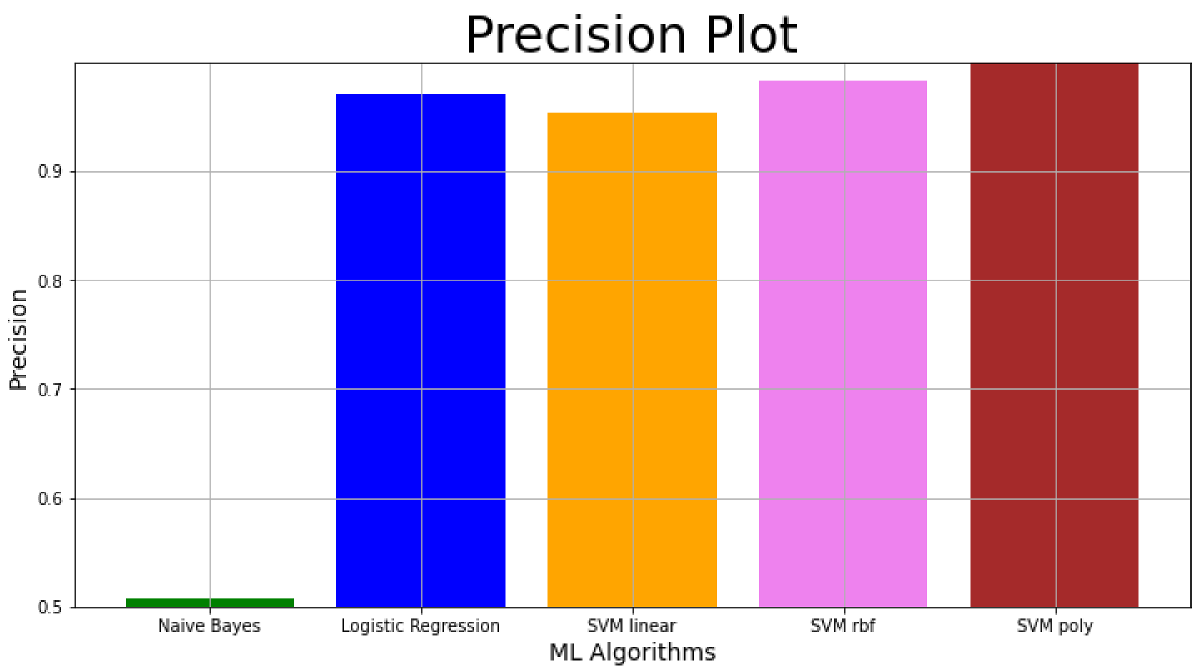

Figure 14: Precision Plot for ML Algorithms using TF-IDF Model

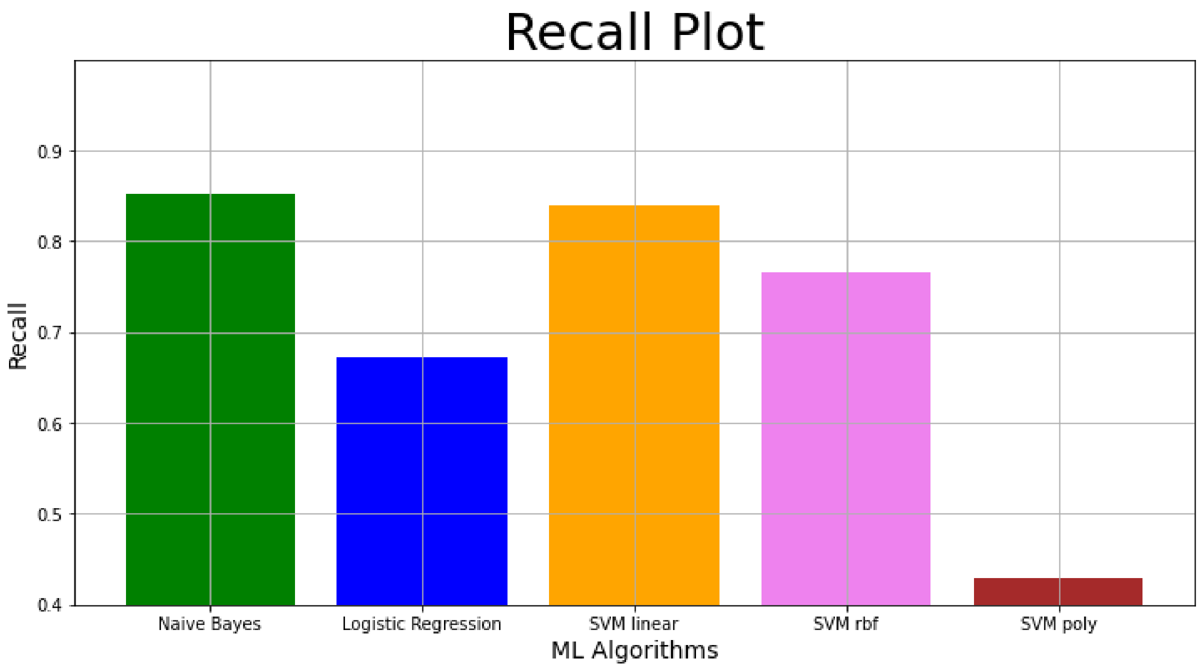

Figure 15: Recall Plot for ML Algorithms using TF-IDF Model

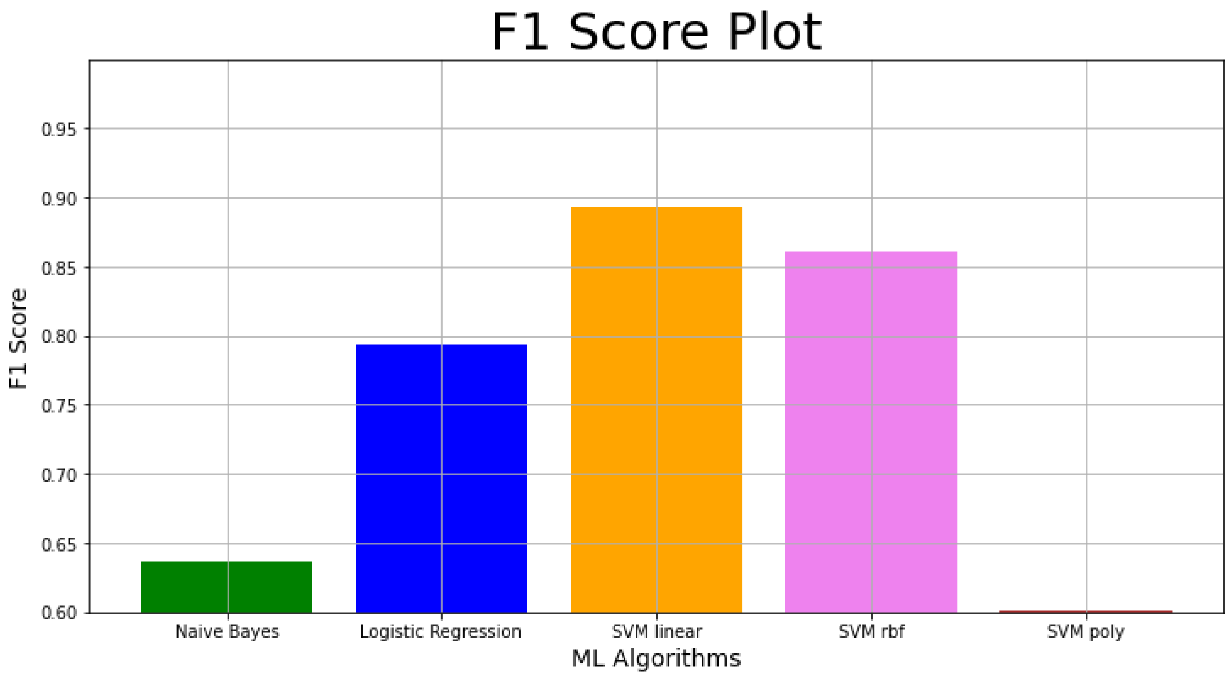

Figure 16: F1-Score Plot for ML Algorithms using TF-IDF Model

Table 4: True Positives, False Positives, True Negatives, and False Negatives result for TF-IDF Model

ML Algorithm	True Positives	False Positives	True Negatives	False Negatives
Naive Bayes	843	22	127	123
Logistic Regression	963	49	100	3

Support Vector Machine (Linear Kernel)	960	24	125	6
Support Vector Machine (RBF Kernel)	964	35	114	2
Support Vector Machine (Poly Kernel)	966	85	64	0

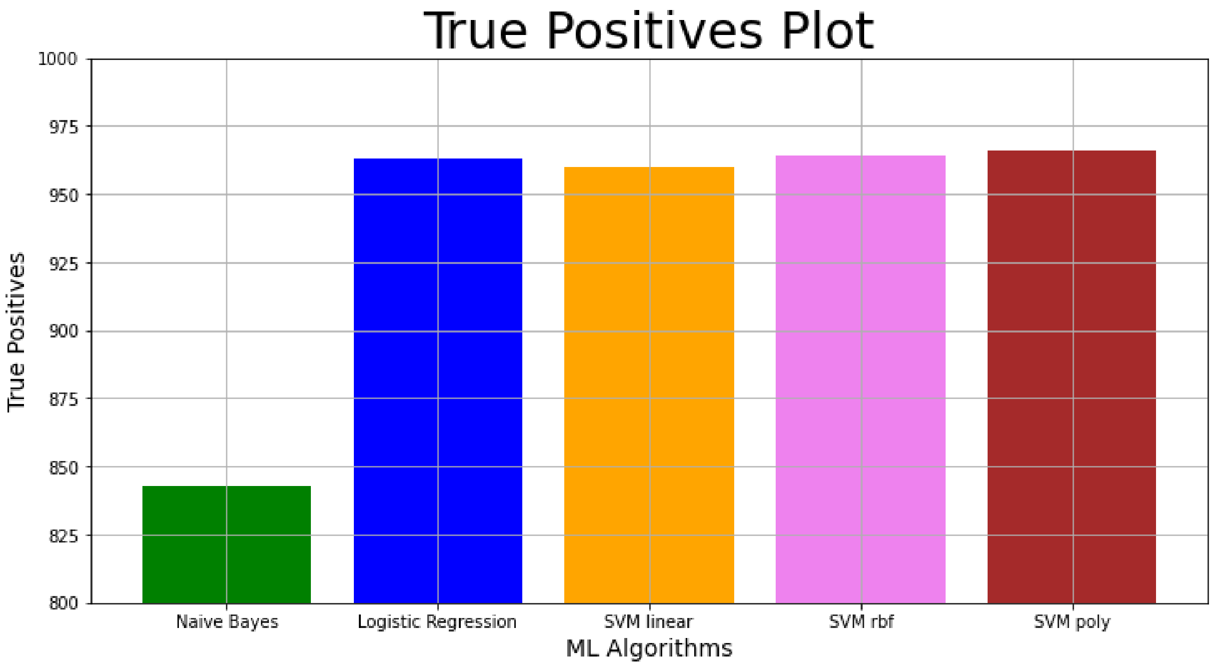

Figure 17: True Positives for ML Algorithms using TF-IDF Model

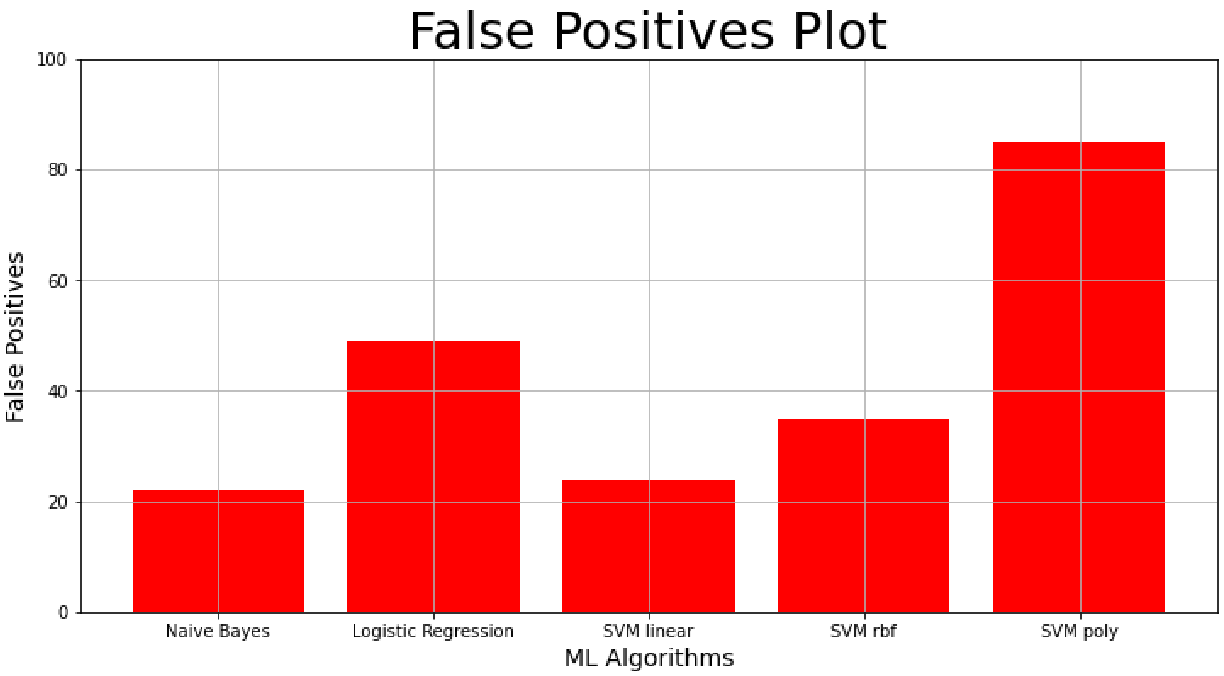

Figure 18: False Positives for ML Algorithms using TF-IDF Model

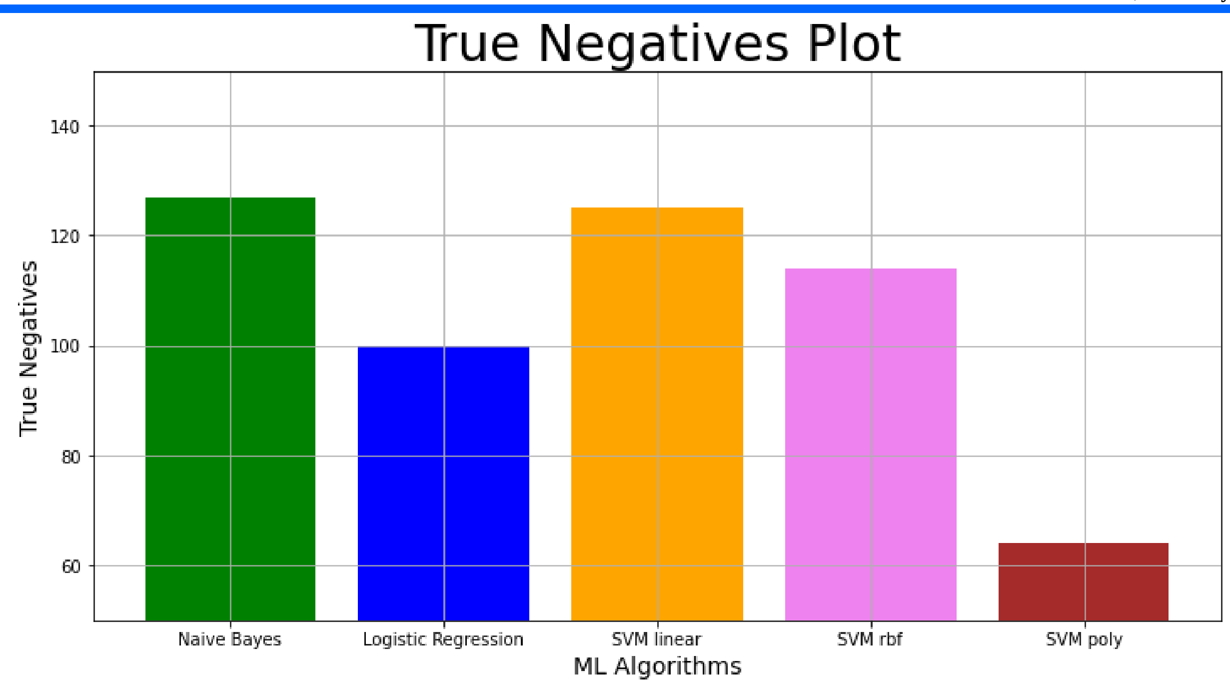

Figure 19: True Negatives for ML Algorithms using TF-IDF Model

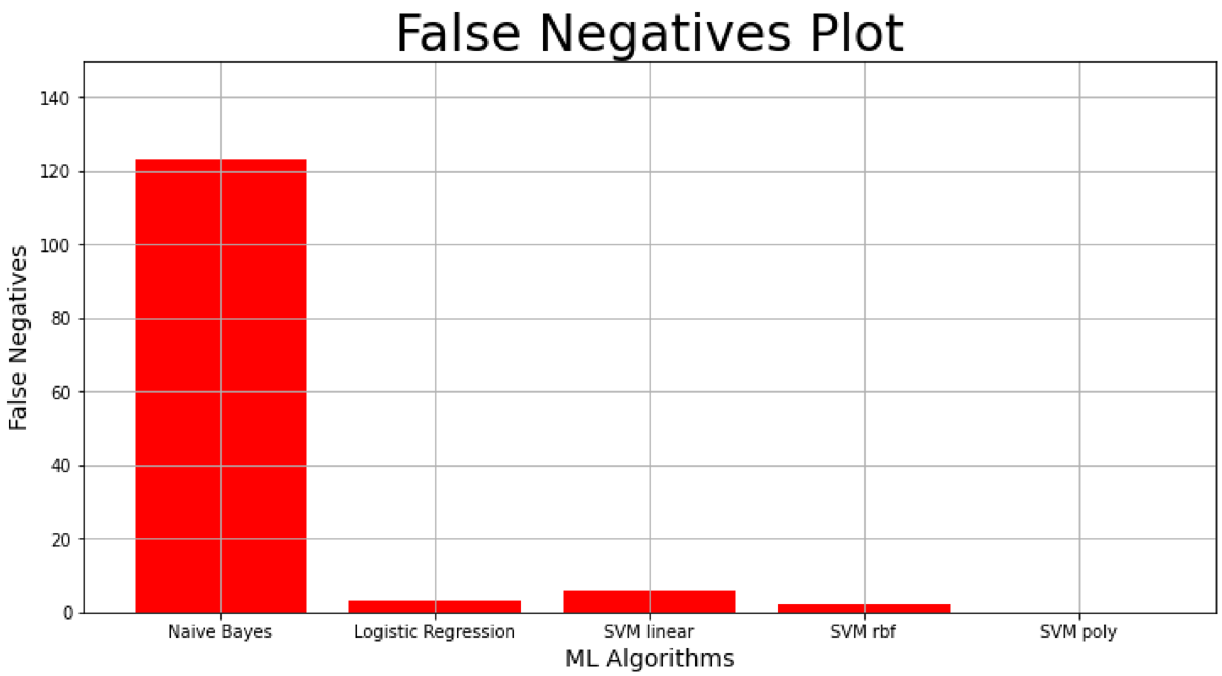

Figure 20: False Negatives for ML Algorithms using TF-IDF Model

4. Conclusion

Machine learning-based spam filter model for SMS is presented. A case study dataset that consists of a large text file that contains SMS messages obtained, pre-processed and used in two Natural Language Processing (NLP) techniques, namely; Bag of Words and Term Frequency Inverse Document Frequency (TF-IDF) to classify the SMS message as ham or spam. Specifically, after pre-processing the dataset, the following machine learning algorithms were used to train the model, namely; Naive Bayes, Support Vector Machine (SVM), and Logistic Regression algorithm. The model was trained using a CPU from Google Colab. The results showed that the machine learning models trained using the Bag of Words model performed better than those trained using TF-IDF model.

References

1. Kim, B. Y., & Lee, J. (2017). Smart devices for older adults managing chronic disease: a scoping review. *JMIR mHealth and uHealth*, *5*(5), e7141.
2. Kirtana, R. N., & Lokeswari, Y. V. (2017, January). An IoT based remote HRV monitoring system for hypertensive patients. In *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)* (pp. 1-6). IEEE.

3. Yasmin, A., Tasneem, S., & Fatema, K. (2015). Effectiveness of digital marketing in the challenging age: An empirical study. *International journal of management science and business administration*, *1*(5), 69-80.
4. Nyeko, J. S., Moya, M., Kabaale, E., & Odongo, J. (2014). Factors influencing the short message service (SMS) mobile banking adoption: a users perspective in the West Nile region in Uganda. *European Journal of Business and Management*.
5. Wei, F., Li, Y., Roy, S., Ou, X., & Zhou, W. (2017, July). Deep ground truth analysis of current android malware. In *International conference on detection of intrusions and malware, and vulnerability assessment* (pp. 252-276). Springer, Cham.
6. Chen, L., Yan, Z., Zhang, W., & Kantola, R. (2015). TruSMS: A trustworthy SMS spam control system based on trust management. *Future Generation Computer Systems*, *49*, 77-93.
7. Döring, N., & Pöschl, S. (2017). Nonverbal cues in mobile phone text messages: The effects of chronemics and proxemics. In *The reconstruction of space and time* (pp. 109-135). Routledge.
8. Zhang, Y., Wang, L., & Duan, Y. (2016). Agricultural information dissemination using ICTs: A review and analysis of information dissemination models in China. *Information processing in agriculture*, *3*(1), 17-29.
9. Bouhnik, D., & Deshen, M. (2014). WhatsApp goes to school: Mobile instant messaging between teachers and students. *Journal of Information Technology Education. Research*, *13*, 217.
10. Park, K., & Kim, H. (2015, August). Encryption Is Not Enough: Inferring user activities on KakaoTalk with traffic analysis. In *International Workshop on Information Security Applications* (pp. 254-265). Springer, Cham.
11. ITU (2019), Measuring digital development: Facts and Figures 2019. Available at https://www.itu.int/myitu/-/media/Publications/2020-Publications/Measuring-digital-development-2019.pdf
12. Ye, H., Malu, M., Oh, U., & Findlater, L. (2014, April). Current and future mobile and wearable device use by people with visual impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3123-3132).
13. Nhavoto, J. A., Grönlund, Å., & Klein, G. O. (2017). Mobile health treatment support intervention for HIV and tuberculosis in Mozambique: Perspectives of patients and healthcare workers. *PloS one*, *12*(4), e0176051.
14. Hilliard, M. E., Hahn, A., Ridge, A. K., Eakin, M. N., & Riekert, K. A. (2014). User preferences and design recommendations for an mHealth app to promote cystic fibrosis self-management. *JMIR mHealth and uHealth*, *2*(4), e3599.
15. Rahman, F. (2019). Trends in Reading Literary Fiction in Print and Cyber Media by Undergraduate Students of Hasanuddin University. *International Journal of Education and Practice*, *7*(2), 66-77.
16. Ling, R., & Lai, C. H. (2016). Microcoordination 2.0: Social coordination in the age of smartphones and messaging apps. *Journal of Communication*, *66*(5), 834-856.
17. Rathbone, A. L., & Prescott, J. (2017). The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. *Journal of medical Internet research*, *19*(8), e7740.
18. Kamal, A. K., Shaikh, Q., Pasha, O., Azam, I., Islam, M., Memon, A. A., ... & Khoja, S. (2015). A randomized controlled behavioral intervention trial to improve medication adherence in adult stroke patients with prescription tailored Short Messaging Service (SMS)-SMS4Stroke study. *BMC neurology*, *15*(1), 1-11.
19. Alzahrani, A. J., & Ghorbani, A. A. (2014, May). SMS mobile botnet detection using a multi-agent system: research in progress. In *Proceedings of the 1st International Workshop on Agents and CyberSecurity* (pp. 1-8).
20. Alzahrani, A. (2016). *An SMS-based mobile botnet detection framework using intelligent agents* (Doctoral dissertation, University of New Brunswick.).
21. Nirbhavane, M., & Prabha, S. (2014). Accident monitoring system using wireless application. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *3*(4), 1532-1535.
22. Pervaiz, F., Nawaz, R. S., Ramzan, M. U., Usmani, M. Z., Mare, S., Heimerl, K., ... & Razaq, L. (2019, July). An assessment of SMS fraud in Pakistan. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies* (pp. 195-205).
23. Xiao, X., Fu, P., Hu, G., Sangaiah, A. K., Zheng, H., & Jiang, Y. (2017). SAIDR: A new dynamic model for SMS-based worm propagation in mobile networks. *IEEE Access*, *5*, 9935-9943.
24. Skudlark, A. (2014). Characterizing SMS spam in a large cellular network via mining victim spam reports.
25. Chen, L., Yan, Z., Zhang, W., & Kantola, R. (2015). TruSMS: A trustworthy SMS spam control system based on trust management. *Future Generation Computer Systems*, *49*, 77-93.
26. Mokri, M. A. E. S., Hamou, R. M., & Amine, A. (2019). A new bio inspired technique based on octopods for spam filtering. *Applied Intelligence*, *49*(9), 3425-3435.
27. Eshmawi, A. A. (2015). *The roving proxy for SMS spam and phishing detection* (Doctoral dissertation, Southern Methodist University).
28. Reaves, B., Scaife, N., Tian, D., Blue, L., Traynor, P., & Butler, K. R. (2016, May). Sending out an SMS: Characterizing the Security of the SMS Ecosystem with Public Gateways. In *2016 IEEE*

Symposium on Security and Privacy (SP) (pp. 339-356). IEEE.
29. Reaves, B., Vargas, L., Scaife, N., Tian, D., Blue, L., Traynor, P., & Butler, K. R. (2018). Characterizing the security of the SMS ecosystem with public gateways. *ACM Transactions on Privacy and Security (TOPS)*, *22*(1), 1-31.
30. Ford, G. S., & Stern, M. L. (2016). Proper Incentives? The Economics of Spam Management by the Mobile Wireless Industry. *The Economics of Spam Management by the Mobile Wireless Industry (May 4, 2016)*.
31. Almossawi, M. (2014). Effectiveness of SMS advertising (a study of young customers in Bahrain). *Global Journal of Management and Business Research*.
32. Nieborg, D. B., & Helmond, A. (2019). The political economy of Facebook's platformization in the mobile ecosystem: Facebook Messenger as a platform instance. *Media, Culture & Society*, *41*(2), 196-218.
33. Vukovic, D. R., & Dujlovic, I. M. (2016, November). Facebook messenger bots and their application for business. In *2016 24th Telecommunications Forum (TELFOR)* (pp. 1-4). IEEE.
34. Yudhana, A., Riadi, I., & Anshori, I. (2018). Analisis Bukti Digital Facebook Messenger Menggunakan Metode Nist. *IT Journal Research and Development*, *3*(1), 13-21.
35. Olasina, G. (2015). Factors influencing the use of m-banking by academics: case study sms-based m-banking. *The African Journal of Information Systems*, *7*(4), 4.
36. Nyakomitta, P. S., & Omollo, V. N. (2014). SMS-Based Alert Notification for Credit Applications Queuing Systems. *International Journal of Innovation and Applied Studies*, *9*(3), 1291.
37. Pentaiah, K. N., & Ker, P. J. (2017). Research Article Development of a Microcontroller-based Portable Surveillance System with User Alert Notification.
38. Sadhasivam, J., Alamelu, M., Radhika, R., Ramya, S., Dharani, K., & Jayavel, S. (2017, November). Enhanced way of securing automated teller machine to track the misusers using secure monitor tracking analysis. In *IOP Conference Series: Materials Science and Engineering* (Vol. 263, No. 4, p. 042032). IOP Publishing.
39. Coull, S. E., & Dyer, K. P. (2014). Traffic analysis of encrypted messaging services: Apple imessage and beyond. *ACM SIGCOMM Computer Communication Review*, *44*(5), 5-11.
40. Coull, S., & Dyer, K. (2014). Privacy failures in encrypted messaging services: Apple imessage and beyond. *arXiv preprint arXiv:1403.1906*.
41. Mark, D., Varma, J., LaMarche, J., Horovitz, A., & Kim, K. (2015). Messaging: Mail, Social and iMessage. In *More iPhone Development with Objective-C* (pp. 319-335). Apress, Berkeley, CA.
42. Choudhary, N., & Jain, A. K. (2017, March). Towards filtering of SMS spam messages using machine learning based technique. In *International Conference on Advanced Informatics for Computing Research* (pp. 18-30). Springer, Singapore.
43. Kumar, S., Gao, X., Welch, I., & Mansoori, M. (2016, March). A machine learning based web spam filtering approach. In *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)* (pp. 973-980). IEEE.
44. Sharma, P., & Bhardwaj, U. (2018). Machine learning based spam e-mail detection. *International Journal of Intelligent Engineering and Systems*, *11*(3), 1-10.
45. Chetty, G., Bui, H., & White, M. (2019, December). Deep learning based spam detection system. In *2019 International Conference on Machine Learning and Data Engineering (iCMLDE)* (pp. 91-96). IEEE.
46. Gheewala, S., & Patel, R. (2018, February). Machine learning based Twitter Spam account detection: a review. In *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 79-84). IEEE.
47. Tandel, S., & Jamadar, A. (2018). Impact of progressive web apps on web app development. *International Journal of Innovative Research in Science, Engineering and Technology*, *7*(9), 9439-9444.
48. Sheppard, D., & Sheppard, D. (2017). *Beginning progressive web app development*. Apress.
49. Nugroho, L. E., Pratama, A. G. H., Mustika, I. W., & Ferdiana, R. (2017, October). Development of monitoring system for smart farming using Progressive Web App. In *2017 9th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 1-5). IEEE.
50. Feng, W., Sun, J., Zhang, L., Cao, C., & Yang, Q. (2016, December). A support vector machine based naive Bayes algorithm for spam filtering. In *2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC)* (pp. 1-8). IEEE.
51. Dong, L., Li, X., & Xie, G. (2014, February). Nonlinear methodologies for identifying seismic event and nuclear explosion using random forest, support vector machine, and naive Bayes classification. In *Abstract and Applied Analysis* (Vol. 2014). Hindawi.
52. Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, *5*(2), 221.
53. Tuhuteru, H., & Iriani, A. (2018). Analisis Sentimen Perusahaan Listrik Negara Cabang Ambon Menggunakan Metode Support Vector Machine dan Naive Bayes Classifier. *Jurnal Informatika*, *3*(03).
54. Feng, J., Xu, H., Mannor, S., & Yan, S. (2014). Robust logistic regression and

classification. *Advances in neural information processing systems*, *27*.

55. De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), 760-772.
56. Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: binary classification for heterogeneous datasets. *SMU Data Science Review*, *1*(3), 9.
57. Mustafa, A., Heppenstall, A., Omrani, H., Saadi, I., Cools, M., & Teller, J. (2018). Modelling built-up expansion and densification with multinomial logistic regression, cellular automata and genetic algorithm. *Computers, Environment and Urban Systems*, *67*, 147-156.
58. De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis*, *26*(4), 417-430.
59. El-Din, D. M. (2016). Enhancement bag-of-words model for solving the challenges of sentiment analysis. *International Journal of Advanced Computer Science and Applications*, *7*(1).
60. González, L. C., Moreno, R., Escalante, H. J., Martínez, F., & Carlos, M. R. (2017). Learning roadway surface disruption patterns using the bag of words representation. *IEEE Transactions on Intelligent Transportation Systems*, *18*(11), 2916-2928.
61. Vetrivel, A., Gerke, M., Kerle, N., & Vosselman, G. (2016). Identification of structurally damaged areas in airborne oblique images using a visual-bag-of-words approach. *Remote sensing*, *8*(3), 231.
62. Rofiqi, M. A., Fauzan, A. C., Agustin, A. P., & Saputra, A. A. (2019). Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query. *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, *1*(2), 58-64.
63. Melita, R. (2018). *Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim)* (Bachelor's thesis, Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta).
64. Kanchana, S., Meenakshi, K., & Ganapathy, V. (2017). Comparison of genre based tamil songs classification using term frequency and inverse document frequency. *Research Journal of Pharmacy and Technology*, *10*(5), 1449.
65. Dadgar, S. M. H., Araghi, M. S., & Farahani, M. M. (2016, March). A novel text mining approach based on TF-IDF and Support Vector Machine for news classification. In *2016 IEEE International Conference on Engineering and Technology (ICETECH)* (pp. 112-116). IEEE.