A proposal of systematic SMS spam detection model using supervised machine learning classifiers

Author

Abstract— SMS (Short Message Service) is a conventional and facile medium for communication around the world. Nowadays each individual in each generation has their handheld mobile device and they are also using SMS services for means of quick communication. On this account, SMS services become a desirable medium for spammers. To diminish this problem, there are many content & non-content-based approaches are being introduced for detecting spam SMS on mobile devices. But they are not as comprehensive as compared with Email spam detection. Hence in this paper, we tried to develop a systematic model using machine learning classifiers to detect SMS spam. We used five supervised machine learning classifiers and a large dataset that contains 11132 data labeled as spam or ham. Then we applied different data preprocess methods and feature extraction on the dataset, to train and test the ML classifiers and evaluate the model in terms of accuracy, precision, recall, f1-score. Final results indicated that Logistic Regression (LR) showed the lowest evaluation results for the model with an accuracy score of 99.28%, precision score of 99.27%, recall score of 95.16%, and F1-score of 97.17%. On the contrary, Random Forest (RF) showed the best classification for the model with an accuracy score of 99.73%, precision score of 1.0%, recall score of 97.92%, and F1-score of 98.95%. The overall procedure is implemented on Jupyter notebook using the python programming language.

Index Terms— SMS, Spam, Ham, detection, Machine learning, evaluation**.**

1 INTRODUCTION

The increased penetration of the use of mobile devices can be attributed to the advancements in telecommunication technologies. The capabilities that the networks offer the users is another factor behind the onward use of mobile devices [1]. Besides, the mobile device market has experienced substantial growth over this period. On a survey, it is shown that a total of 4.78 million mobile phones have shipped over the year 2020, which shows an increment of about 6% over the year [2]. In this era of modern technology, the use of free services has increased which is available on the internet like- Messenger, iMessenger, WhatsApp, Viber, etc. Despite all this availability and modern communication medium, SMS is still a strong and ingrown medium on mobile devices communication. According to Statista, the total number of mobile devices is expected to reach 17.72 billion by 2024 worldwide [3].

Many businesses are increasingly using SMS services. For example, the corporate offices, banking, financial services, insurance industries, and even different online communication mediums have implemented SMS services to a broad degree for verifications, notifications, and OTPs(one-time-pins). Also, nowadays Internet of Things (IoT) devices are becoming very popular and many of those devices use SMS services only because of their high availability, large coverage, high deliverability, and importantly most reliability offered. In both home and workplace, SMS services are used by these IoT devices to do things like delivering time-sensitive notification on changes in server status so that it can provide continuous information at a variety of sensors, as for temperature sensors, power sensors, water sensors, and humidity sensors and so on. It is expected that at the time of 2025, globally there will be 25 billion IoT connections.

Short Message Service (SMS) can be referred to as solicited and unsolicited or spam text. Whether spam is in the form of email or SMS, the danger is equal. In the world, the total number of mobile device users is around 15.96 billion at the time 2020 [4]. Nowadays users have their confidential and personal information like credit card numbers, contact lists, passwords, banking information, and so on stored in those devices. And by using spam SMS hackers can easily access this information from users' smartphone devices without the knowledge of the user. Which leads to various cybercrime activities. Nowadays with the advancement of technology, SMS spam also increases drastically which is an annoyance for users and users also face critical data loss.

2 RELATED WORKS

Even though in the last period there are few works done on spam SMS detection but there are only some observable works available on this topic. Tran Ho-Seok K. and Sung-Ryul K. proposed a method using a KNN classifier along with a graph-based algorithm, where the model detects spam SMS with high accuracy and shows small process time [5]. Another method proposed by Rish I., where they proposed a method of low entropy feature distribution which yields the best performance for the Naive Bayes classifier [6].

An article derived by M. Bassiouni, M. Ali, and E. A. El-Dahshan, where they used 10 different classifiers on a one benchmark dataset, where they evaluate the best classifier that gave the best result [7]. By using open source WEKA data mining tool Yang and S. Elfayoumy explore the performance analysis from different classifiers, from where they identified the superlative classifier for the classification of spam [8]. By using content-based spam detection methods S. J. Delany, M. Buckley, and D.Greene proposed a method and evaluated them accordingly [9]. A content-based spam filter method is proposed by Tiago A. Almeida, José María Gómez, Akebo Yamakami, where they used a non-encoded, real, public SMS spam collection and then utilized it to find out the best model with the highest accuracy for spam detection [10]. A paper using a content-based method approached by El-Sayed, M. El-Alfy, ennobles mobile device communication by detecting SMS spam with the highest accuracy [11]. By implementing ML classifiers, Dr. Dipak R. Kawade, Dr. Kavita S. Oza, provides a content-based SMS spam detection model which gives an accuracy rate of 98% [12]. There is another effective method for SMS spam detection which is called Latent-Content based feature, and that is proposed by Amir K. and Lina Zh. [13].

Dendritic Cell Algorithm (DCA) is a hybrid method of SMS spam detection that is proposed by Julie Greensmith1 and Uwe Aickelin, which shows good performance when it's applied to a real-time dataset [14]. A paper which is proposed by Housh and Shirani-Mehr, applied machine learning classifiers to solve SMS spam and they got a model with an accuracy of 97.64% [15]. Another model proposed by Mehul Gupta, Aditya Bakliwal, Shubhangi Agarwal, Pulkit Mehndiratta, in which they implied various ML classifiers for SMS spam filtering in python using Google collabo. [16]. A survey provided by Hedieh Sajedi, Golazin Zarghami Parast, Fatemeh Akbari, where they collected all of the existing Learning-based methods for spam detection and gives a structural overview. They select almost 44 articles, found 28 extraction methods, and among them 15 classifiers have been compared their Accuracy, weakness, error, strengths in spam filtering. Among all of them, DCA the large cellular network model, and KNN showed the best result in SMS spam detection. And one of the classifiers showed an accuracy rate of 98.63% [17]. Another spam detection model showed up accuracy of around 95.45% using the Random Forest classifier [7].

There are some other distinct methods used in spam detection. A deep learning-based model like Convolutional Neural Network (CNN) and Long short-term memory (LSTM) was proposed by Healy, M., Delany, S.J., Zamolotskikh, A., which introduce semantic information by using knowledge bases as like-WordNet and ConceptNet [18]. Amani Alzahrani and Danda B.Rawat also proposed a model for SMS spam detection using neural networks, where they achieved 98% accuracy with the lowest error rate [19].

3 METHODOLOGY

In this section, we'll discuss the research method and used materials of this research.

3.1 Datasets

We collected data from several online sources. We collected a number of 5574 data contains both ham and spam text from a public online research source. NUS (National University of Singapore) creates a dataset that contain 3375 ham SMS, Grumbletext Website which contain 425 spam SMS, Caroline Tag's Ph.D. Theses that contains 450 SMS which are ham and SMS Spam Corpus v.0.1 contains 322 spam SMS and 1002 ham SMS. To sum up, they got 747 spam SMS and 4827 ham SMS. UCI Machine learning repository hosted this corpus and it is also publicly available in raw format at [20]. Another dataset was collected from Kaggle contains 5156 data in which 87% are ham and 13% are spam text [21]. Then we combined these collected data in a CSV (Comma-Separated Values) file. Finally, we gathered a total of 11132 data where 9637 (86.6%) are ham or solicited SMS and 1494 (13.4%) are spam or unsolicited SMS.

3.2 Model Procedure

We divided our system into some sub-systems and each sub-system is presented in the flow diagram of Figure 1. The whole procedure is implemented on jupyter notebook by using the python programming language.

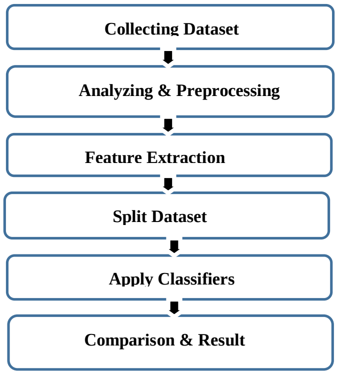
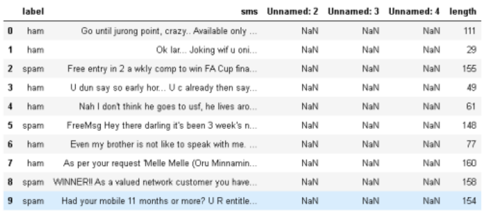


Fig. 1. Flow diagram of our proposed method.

3.2.1 Preparing the dataset

To prepare the datasets, we have used CSV (Comma Separated Values) file. The file contains one SMS per line. The file contains two columns, one is titled as a label which contains two types of labeling (ham and spam), another is titled as sms which contains raw text messages. Now we imported the dataset in juypter notebook using pandas.



	label	sms	Unnamed: 2	Unnamed: 3	Unnamed: 4	length
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN	111
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN	29
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN	155
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN	49
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN	61
5	spam	FreeMsg Hey there darling it's been 3 week's n...	NaN	NaN	NaN	148
6	ham	Even my brother is not like to speak with me. ...	NaN	NaN	NaN	77
7	ham	As per your request 'Melle Melle (Oru Minnamin...	NaN	NaN	NaN	160
8	spam	WINNER!! As a valued network customer you have...	NaN	NaN	NaN	158
9	spam	Had your mobile 11 months or more? U R entitle...	NaN	NaN	NaN	154

Fig. 2. Importing Dataset with all columns

3.2.2 Data Analyzing & Preprocessing

we dropped all the unwanted columns from the dataset. Before implementation, data analysis is mandatory for getting a clean and accurate model. In figure 3, we can see the histogram, which shows the length of each sms.

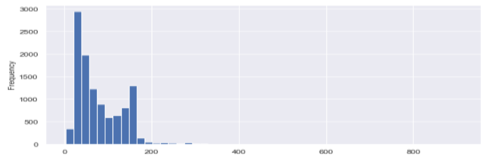


Fig. 3. Showing length of each sms.

Theoretically, SMS with large lengths are detected as spam. On the other hand, SMS with shorter lengths are categorized as ham (legitimate SMS).

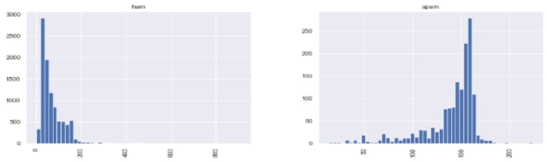


Fig. 4. Showing length of ham and spam SMS.

Fig. 4 proves that spam SMS got larger length than ham. For easier analysis, we converted the labeling in binary format, as "o" as ham and "1" as spam.

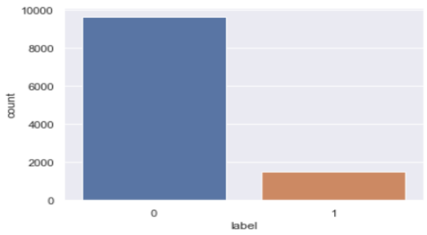


Fig. 5. Ham and spam words per text.

We also applied several preprocessing approaches to clean the data. We removed the stop words and also performed stemming and lemmatization on the data.

3.2.3 Feature Extraction

Feature extraction is used to get the best characteristics. It is required to categorize the SMS and also to remove the noisy features. Moreover, it also helps to increase the accuracy, avoids overfitting, simplifying the overall calculation. There are varieties of feature extraction methods or processes are available. Since our dataset consists long text file, in which each text line corresponds to a text message, we used the Bag of Words(BOW) approach.

Bag of Words (BOW) Approach: Our dataset contains 11132 data of which 9637 are ham and 1494 are spam SMS. As we know that machine learning classifiers need numerical data as input for performing classification, we needed an accurate process to convert the text messages into numerical data for better implementation. We used the BOW feature extraction approach to helps us to get that outcome. There are four steps followed in the BOW approach. They are- converting all strings in their lowercase, remove all punctuations, Tokenize the word, count word frequency by importing count from the counter.

CountVectorizer: Firstly, to clean our data from scratch, we need to implement CountVectorizer(). This cleaning process indicates convert all the stored data in their lowercase form and then punctuation marks should be removed from all. The parameter used in CountVectorizer is:

lowercase = True

There is a default value of True for these lowercase parameters which is used to convert the text in its lowercase.

Token pattern = (?u)\b\w\w+\b

3.2.4 Visualizing Data

This is the analytical phase of the development of the model solution. The analysis is formed following the number of spam and ham and also the number of word counts for each SMS text. From the counts of the SMS text, the mathematical analysis shows that there are 86.6% ham SMS and 13.4% spam SMS. The pie chart in Figure 6 demonstrates that.

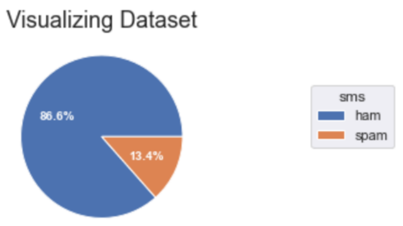


Fig. 6. Visualization of the dataset.

3.2.5 Split Dataset into Train and Test

Using sklearn model selection we split the dataset into train and test for further implementation. We used the training data to train the ML classifiers and the testing data to evaluate the performance of the model.

4 Performance Evaluation and Discussion

4.1 Performance Metrics

To get expected results, it is mandatory to choose performance metrics correctly. This helps to analyze and evaluate the model accurately. To evaluate the implemented classifiers we have taken four metrics. which are as follow:

4.1.1 Accuracy

The correct prediction made by the classifiers is referred to as accuracy. In other words, true condition or quality also indicates an accuracy of a model. Accuracy can be written using (1).

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

4.1.2 Precision

The number of text messages which are truly classified as spam can be defined as precision. Precision can be presented as the ratio of true positives (text messages are being classified as true spam) to all the positives predictions. Precision can be written using (2).

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (2)$$

4.1.3 Recall (Sensitivity)

Recall or Sensitivity is defined as the proportion of text messages that truly indicated spam which was classified by us as spam. Recall can be written using (3).

$$\text{Recall(Sensitivity)} = \frac{TP}{(TP+FN)} \quad (3)$$

4.1.4 F1 Score

In many situations, precision and recall can be combined to get a new metric which is known as the f1-score. The range of this score can be between 0 and 1, and 1 is known as the best possible f1 score.

4.2 Evaluating Classifiers

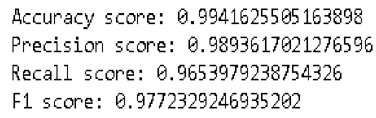
In this section, we will evaluate our classifiers with those four metrics below.

4.2.1 Multinomial Naive Bayes (MNB)

The Naive Bayes classifier is used to build a probabilistic model for SMS spam detection. Naive Bayes proceeds the formula is shown in (4).

$$P(a|b) = \frac{P(a) \times P(a|b)}{P(b)} \quad (4)$$

MNB yielded an accuracy of 99.42%. It is precision and recall scores are 98.93% and 96.54%, F1-score is 97.72%. The results are shown in Figure 7.

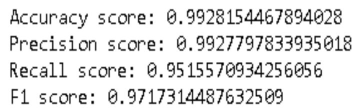


```
Accuracy score: 0.9941625505163898
Precision score: 0.9893617021276596
Recall score: 0.9653979238754326
F1 score: 0.9772329246935202
```

Fig. 7. Performance measures score of MNB.

4.2.2 Logistic Regression(LR)

Logistic regression is highly considered for binary classification. LR yielded an accuracy of 99.28%. It is precision and recall scores are 99.28% and 95.16%, F1-score is 97.17%. The results are shown in Figure 8.

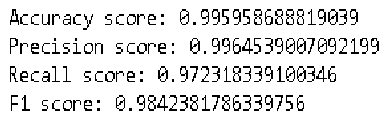


```
Accuracy score: 0.9928154467894028
Precision score: 0.9927797833935018
Recall score: 0.9515570934256056
F1 score: 0.9717314487632509
```

Fig. 8. Performance measures score of LR.

4.2.3 Support Vector Machine(SVM)

Support vector machine is used to analyze data and recognize patterns. SVM yielded an accuracy of 99.60%. It is precision and recall scores are 99.65% and 97.23%, F1-score is 98.42%. The results are shown in Figure 9.

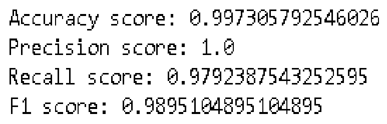


```
Accuracy score: 0.995958688819039
Precision score: 0.9964539007092199
Recall score: 0.972318339100346
F1 score: 0.9842381786339756
```

Fig. 9. Performance measures score of SVM.

4.2.4 Random Forest

Random Forest is a supervised learning method. It is an ensemble method since it is built up by using several decision trees. Random forest yielded an accuracy of 99.73%. It is precision and recall scores are 1.0%, and 97.92%, F1-score is 98.95%. The results are shown in Figure 10.

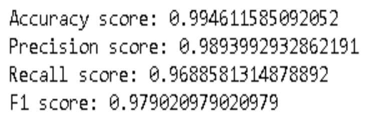


```
Accuracy score: 0.997305792546026
Precision score: 1.0
Recall score: 0.9792387543252595
F1 score: 0.9895104895104895
```

Fig. 10. Performance measures score of RF.

4.2.5 Decision Tree

Decision Tree is known as a structured non-parametric approach for problem-solving. Decision Tree yielded an accuracy of 99.46%. It is precision and recall scores are 98.94% and 96.89%, F1-score is 97.90%. The results are shown in Figure 11.

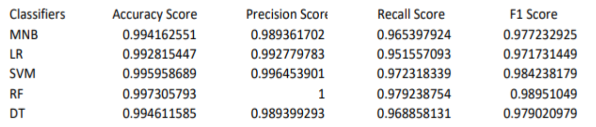


```
Accuracy score: 0.994611585092052
Precision score: 0.9893992932862191
Recall score: 0.9688581314878892
F1 score: 0.979020979020979
```

Fig. 11. Performance measures score of DT.

4.2.6 Summary of Classifiers

The detailed results of all the classifiers are shown in figure 12.



Classifiers	Accuracy Score	Precision Score	Recall Score	F1 Score
MNB	0.994162551	0.989361702	0.965397924	0.977232925
LR	0.992815447	0.992779783	0.951557093	0.971731449
SVM	0.995958689	0.996453901	0.972318339	0.984238179
RF	0.997305793	1	0.979238754	0.98951049
DT	0.994611585	0.989399293	0.968858131	0.979020979

Fig. 12. Showing the detailed results for the proposed model.

4.3 Plot ROC & Compare AUROC

The Receiver Operating Characteristic(ROC) curve is known as a graph that is used to evaluate the performances of the classification models in all of its thresholds. There are two important parameters that exists in this curve. They are-

True Positive Rate or TPR can be written using (5).

$$\text{True Positive Rate} = \frac{\text{TP}}{(\text{TP+FN})} \qquad (5)$$

False Positive Rate or FPR can be written using (6).

$$\text{False Positive Rate} = \frac{\text{FP}}{(\text{FP+TN})} \qquad (6)$$

The area under the ROC Curve or AUC is used to determine the whole 2D (two dimensional) area underneath the whole ROC curve from value (0,0) to (1,1). The performance of an AUROC is known as sub-optimal performance if AUROC is less than 0.70, good performance if AUROC is in between 0.70 - 0.80, and excellent performance if AUROC is greater than 0.8. And it is referred to as a perfect classifier if AUROC is 1.0. In Figure 13 as we can see that, random forest classifiers AUROC score is 1.0, it indicates that RF is a perfect classifier.

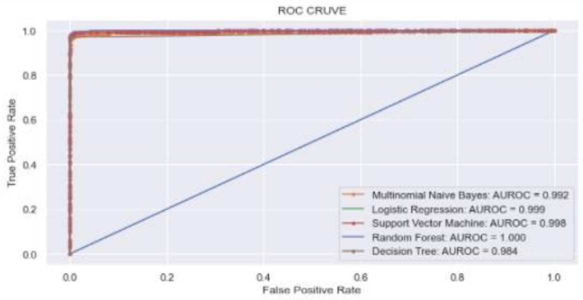


Fig. 13. Plotting ROC & Compare AUROC for multiclass classifiers.

4.4 Performance Metrics Curve

In this part, we will individually take scores and compare the results of four metrics for the five classifiers. In the below curves, different classifiers are put on the x-axis and Performance metrics are on the y-axis.

4.4.1 Plotting Accuracy Score

Random Forest got the highest accuracy score. The accuracy of Random Forest (RF) is 99.73%.

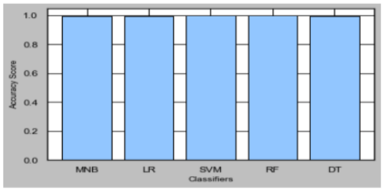


Fig. 14. Accuracy scores of the classifiers.

4.4.2 Plotting Precision Score

In the precision curve, RF got the highest precision score among all of the classifiers. The Precision score of RF is 100%.

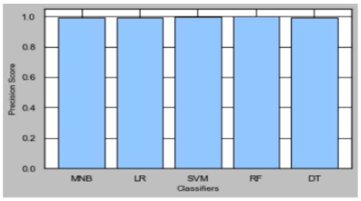


Fig. 15. Precision scores of the classifiers.

4.4.3. Plotting Recall Score

In the recall(sensitivity) curve, both RF and SVM got the highest precision score among all of the classifiers. But between the two of them, RF got the highest recall score that is 97.92%.

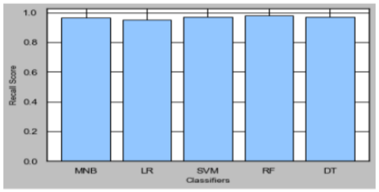


Figure 16: Recall scores of the classifiers.

4.4.4 Plotting F1 Score

RF got the highest F1-score among all of the classifiers. F1-score of Random Forest is 98.95%.

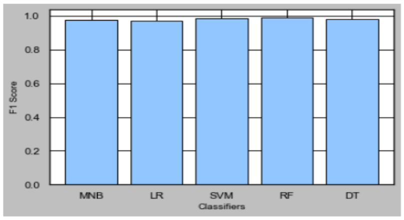


Fig. 17. F1-scores of the classifiers.

5 CONCLUSION AND FUTURE SCOPE

Machine learning is the most popular technique used in the classification of messages into spam or ham. Its successful use in producing email spam classification makes it a viable option for the classification for mobile or SMS spam detection. It is well known in the sector of machine learning that building a 100% accurate model is quite tough. Again due to the lack of combined and organized datasets and the size and structure of SMS, it is difficult to get an accurate model for SMS spam detection. Here, we proposed a model where we used five supervised machine learning classifiers and a large dataset (11132). Finally, we got our desired classifiers with the highest accuracy which is Random Forest (RF) with 99.73% accuracy. In this research, we tried to implement supervised machine learning (SML) for our classification. In our model, we found that Random Forest has the highest accuracy score 99.73%. Also, RF got a 1.0 AUROC score which indicated that RF is the perfect classifier for SMS spam detection. In the future, this research can be used in real-world problem solutions for the detection of spam SMS.

REFERENCES

[1] D. B. R. a. K. Z. Ghafoor, Smart Cities Cybersecurity and Privacy, London: Elsevier Press, November 2018.

[2] D. Nagel, "Smart Phone Growth Accelerates Worldwide," 29 07 2013. [Online].Available: https://thejournal.com/articles/2013/07/29/smart-phone-growth-accelerates-worldwide.aspx. [Accessed 29 7 2013].

[3] S. O'Dea, "Statista," 18 12 2020. [Online]. Available: https://www.statista.com/statistics/245501/multiple-mobile-device-ownership-worldwide/. [Accessed 01 01 2020].

[4] S. R. Department, " Statista," 23 11 2016. [Online]. Available: https://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide/. [Accessed 23 16 2016].

[5] H.-S. K. a. S.-R. K. Tran H., "Graph-based KNN Algorithm for Spam SMS Detection," *Journal of Universal Computer Science,* vol. 19, no. 16, pp. 2404-2419, 2013.(submitted for publication)

[6] Rish I,T.J. Watson Research Centre "An empirical study of the Naive Bayes Classifiers", unpublished. (Unpublished)

[7] M. A. a. E. A. E.-D. M. Bassiouni, "Ham and spam E-mails classification using machine learning techniques," *Journal of Applied Security Research,* vol. 13, no. 3, p. 315– 331, 2018.(submitted for publication)

[8] Y. a. S. Elfayoumy, "Anti-Spam Filtering Using Neural Networks and Bayesian Classifiers," in *Proceedings of the 2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, Jacksonville, FL, USA, 2007. (submitted for publication)

[9] M. B. a. D. G. S. J. Delany, "SMS spam filtering: Methods and data," *Expert Systems with Applications,* vol. 39, pp. 9899-9908, 2012. (submitted for publication)

[10] J .M. G. A. Y. Tiago A. Almeida, "Contributions to the Study of SMS Spam Filtering: New Collection and Results," *Proceedings of the 11th ACM symposium on Document engineering,* pp. 259--262, 2011. (submitted for publication)

[11] E.- S. M. E.-A. Ali A. Al-Hasan, "Dendritic Cell Algorithm for Mobile Phone Spam Filtering," *6th International Conference on Ambient Systems, Networks and Technologies, ANT 2015,* pp. 244-251, 2015. (submitted for publication)

[12] D.K. S. O. Dr. Dipak R. Kawade, "CONTENT-BASED SMS SPAM FILTERING USING MACHINE LEARNING TECHNIQUE," *International Journal of Computer Engineering and Applications,* vol. 12, no. 4, pp. 2321-3469, 2018. (submitted for publication)

[13] A. K. a. L. Zh., "Improving Static SMS Spam Detection by Using New Content-based Features," in *20th Americas Conference on Information Systems (AMCIS)*, 2014. (submitted for publication)

[14] G. a. U. Aickelin, "The Deterministic Dendritic Cell Algorithm," *artificial Immune Systems,* vol. 5132, pp. 291-302, 2008. (submitted for publication)

[15] H. a. Shirani-Mehr, "SMS Spam Detection using Machine Learning Approach," Semantic Scholar, 2013, unpublished. (Unpublished)

[16] A. B. S. A. P. M. Mehul Gupta, "A Comparative Study of Spam SMS Detection using Machine Learning Classifiers," in *2018 Eleventh International Conference on Contemporary Computing (IC3)*, Noida, India, 2018. (submitted for publication)

[17] G. Z. P. F. A. Hedieh Sajedi, "SMS Spam Filtering Using Machine Learning Techniques: A Survey.," *Machine Learning Research,* vol. 1, no. 1, pp. 1-14, 2016. (submitted for publication)

[18] M. D. S. Z. Healy, "An assessment of case base reasoning for short text message classification," in *Proceedings of the 15th. Irish Conference on Artificial Intelligence and Cognitive Sciences (AICS'04)*, 2004. (submitted for publication)

[19] A. A. a. D. B.Rawat, "A Comparative Study of Machine Learning Algorithms for Spam SMS Detection," in *Data Science & Cybersecurity Cente , Howard University*, Washington DC 20059, USA. (submitted for publication)

[20] T. A. Almeida, "UCI Machine Learning Repository," 22 06 2012. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection.

[21] V. Chutke, "Kaggle," 11 03 2017. [Online]. Available: https://www.kaggle.com/vivekchutke/spam-ham-sms-dataset.