Spam SMS Detection using Text Classification and Topic Modeling

1$^{\text{st}}$ Vrushabh Wadnere
ENGL.681 Natural Language Processing
Rochester Institute of Technology
Rochester, United States of America
vw1937@rit.edu

***Abstract*—In recent years, there is a drastic increase in scams via mobile short message services (SMS) which are indented to exploit confidential data and financial data and damage the devices of recipients of such messages. Since mobile SMS became an essential part of daily communication in our life. These SMS scams have annoyed mobile subscribers, and such messages are sometimes targeted to categories of subscribers who are more vulnerable to such unsolicited messages. Thus, it became imperative to develop an approach to identify such scam SMS from other messages to provide security to mobile service subscribers. In this paper, we propose a method for detecting spam SMS using various supervised machine-learning techniques and extracting more information from the messages using topic modeling. We tested and compared the performance of our model on various setups. [2]The proposed spam and ham SMS detection system was able to classify spam and ham SMS with an accuracy of 98%.**

***Index Terms*—Spam and Ham SMS detection, Supervised machine learning techniques, Topic modeling.**

I. Introduction

E-communication (email, SMS, and online chat apps) for information sharing has become an essential part of daily life in today's world. Because many organizations and businesses use e-mail and SMS messages to advertise their products, services, and special offers.[2] Spammers can now target mobile subscribers by using third-party automated SMS or E-mail services to send mass spam SMS over an operating network with the intent of stealing users' confidential information in any form, including personal and financial information. This has become a major issue in Western countries, where spam SMS are targeted to specific user groups such as older age groups, teenagers, international students, and others.[5] Spammers can also use SMS spam to infect systems with spyware. To protect mobile network subscribers from such SMS spam, a spam SMS detection system that can assist mobile networks and their users in identifying such messages is required. Thus, our proposed approach for automatically detecting spam SMS using classification and topic modeling techniques can assist mobile networks in protecting their subscribers from spam SMS. This project employs various supervised classification techniques for automatically identifying spam and ham messages by analyzing the most common topics on which such SMS spam is based using topic modeling techniques. In this paper, we have provided a detailed description of the various stages of model development. The organization of the paper is as follows; section 2 covers data collection, data preprocessing, and exploratory data analysis while section 3 covers methodology for spam detection and topic modeling system. The results are discussed in section 4 for each topic. Furthermore, sections 5, 6, 7 and 8 discuss the limitation, future Scope, ethics and conclusion of our research.

II. Data

[2]The dataset used for the spam and ham SMS detection system was prepared from UCI Machine Learning Repository and accessed from Kaggle. The dataset contains a total of 5574 message samples i.e.,724 spam messages and 4850 ham messages. These messages were collected from the mobile phone spam research-related area from students attending the university.

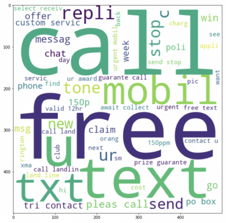


Fig. 1. Word Cloud for Spam messages.

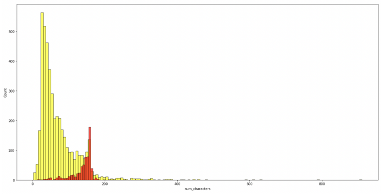


Fig. 2. Character count comparison between Spam and Ham messages.

A. Exploratory Data Analysis

Exploratory data analysis is used in this step to gain insight from the SMS spam collection data set. We discovered 403 duplicates and 0 missing values in our data. [2]Furthermore, we analyzed the text data in terms of the number of characters, words, and sentences to generate new features that can be used to understand the structure of ham and spam messages in the dataset. We discovered that the mean number of characters, words, and sentences in ham messages is 70.90, 17.26, and 1.82, respectively, while the distribution of the same in spam messages is 137.70, 27.76, and 2.98, respectively, allowing us to conclude that spam messages are intended to be more descriptive to capture the attention of the recipients as shown in Fig.2. Furthermore, the creation of a spam message word cloud revealed commonly used words such as text, free, new, send, tone, and guarantee as shown in Fig.1. On the other hand, we discovered that causal vocabulary words like go, come, love, and around are frequently used in ham messages.

B. Data Preprocessing

To improve model efficiency, data must be preprocessed before being used to build models.[6] In this step, text preprocessing is applied to the data set to make it more consistent; text data is converted into lowercase, special characters and numbers are removed from the text, stop-words are removed to reduce low-level information from the data, and stemming is applied to the dataset using PoterStemmer to normalize text data.

III. Methodology

A. Spam and Ham Detection System

Fig. 3. depicts the proposed flow of our model for classifying spam and ham messages. Once the data has been processed, it is ready for model construction. When dealing with textual data, it is critical to convert it into word vector embeddings, which are then passed through the TF-IDF vectorizer and Count Vectorizer before splitting the data into training and testing sets with an 80:20 ratio. In the following step, various supervised machine learning algorithms are used to accurately classify spam and ham messages: [1,5]Gaussian Nave Bayes, Multinomial Nave Bayes, Bernoulli Nave Bayes, Support Vector Classifier, Random Forest, Logistic Regression, Decision Tree, AdaBoost Classifier, Bagging Classifier, Extra Tree Classifier, Gradient Boosting Classifier, Xgboot Classifier, and their accuracies in different settings are compared to identify the best classification technique. [7,8]The model's efficiency is tested in four settings: i) using a count vectorizer, ii) TF-IDF vectorizer with a max feature=3000 limit, iii) Minmax Scaler for feature normalization, and iv) training models with oversampled data using the Synthetic Minority Oversampling technique (SMOTE) to solve the class imbalance.[2] Finally, to improve the overall accuracy of the proposed model ensemble technique, the top three classification models with the highest accuracy and precision score are chosen, which will contribute to the Voting and Stacking classifier for predicting spam and ham messages. In the results section, we will go over model performance in depth.

B. Topic Modeling

Topic modeling is used to analyze the contents of text messages to gain more insight. [9]Since more meaning can be derived from a single word, general text preprocessing is required to optimize text; stop-word removal, tokenization, text lemmatization, punctuation removal, and special character removal are all required. The Latent Dirichlet Allocation (LDA) model is then used to fit data, with each message represented as a topic weight vector. Furthermore, the propensity score is used to categorize the messages' topics.

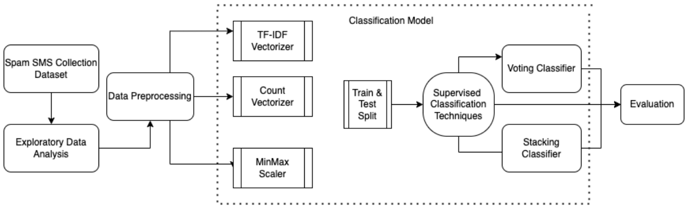
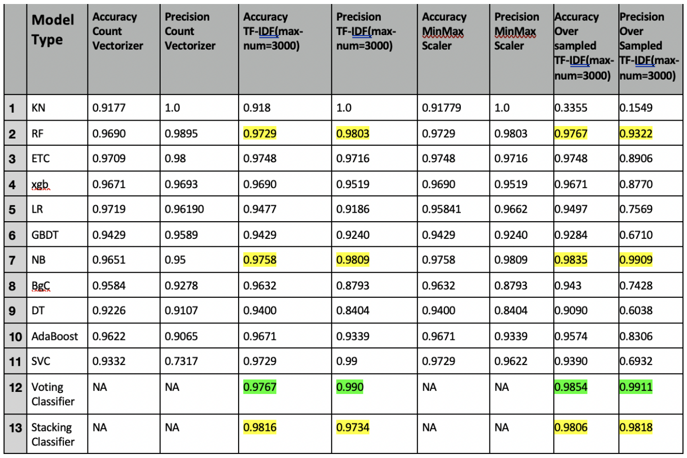


Fig. 3. Architecture Diagram.



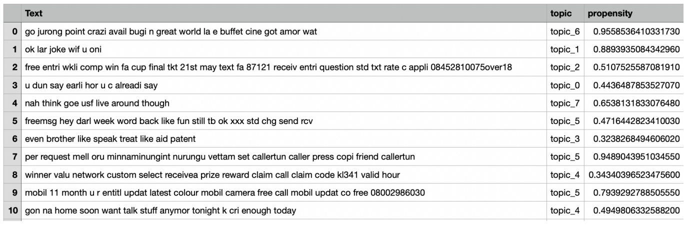
	Model Type	Accuracy Count Vectorizer	Precision Count Vectorizer	Accuracy TF-IDF(max-num=3000)	Precision TF-IDF(max-num=3000)	Accuracy MinMax Scaler	Precision MinMax Scaler	Accuracy Over sampled TF-IDF(max-num=3000)	Precision Over Sampled TF-IDF(max-num=3000)
1	KN	0.9177	1.0	0.918	1.0	0.91779	1.0	0.3355	0.1549
2	RF	0.9690	0.9895	0.9729	0.9803	0.9729	0.9803	0.9767	0.9322
3	ETC	0.9709	0.98	0.9748	0.9716	0.9748	0.9716	0.9748	0.8906
4	xgb	0.9671	0.9693	0.9690	0.9519	0.9690	0.9519	0.9671	0.8770
5	LR	0.9719	0.96190	0.9477	0.9186	0.95841	0.9662	0.9497	0.7569
6	GBDT	0.9429	0.9589	0.9429	0.9240	0.9429	0.9240	0.9284	0.6710
7	NB	0.9651	0.95	0.9758	0.9809	0.9758	0.9809	0.9835	0.9909
8	BgC	0.9584	0.9278	0.9632	0.8793	0.9632	0.8793	0.943	0.7428
9	DT	0.9226	0.9107	0.9400	0.8404	0.9400	0.8404	0.9090	0.6038
10	AdaBoost	0.9622	0.9065	0.9671	0.9339	0.9671	0.9339	0.9574	0.8306
11	SVC	0.9332	0.7317	0.9729	0.99	0.9729	0.9622	0.9390	0.6932
12	Voting Classifier	NA	NA	0.9767	0.990	NA	NA	0.9854	0.9911
13	Stacking Classifier	NA	NA	0.9816	0.9734	NA	NA	0.9806	0.9818

Fig. 4. Model Comparison.

IV. Results

A. Spam and Ham Detection System

Our proposed model demonstrates voting classifier(Random Forest classifier, BernoulliNB, and Extra Tree classifier) provides the best results for spam detection with an accuracy of 98. 49% and F1-score of 93.72%. The proposed model performed better when applying the following setup; balance training dataset and TF-IDF(max-fetures=3000) compared to setup; unbalanced training dataset and TF-IDF(max-fetures=3000) as shown in Fig. 4 . The Naïve Bayes classifier outperformed the Stacking classifier(Random Forest classifier, BernoulliNB, and Extra Tree classifier) providing an accuracy = 98.35%and precision score =0.99 whereas the Stacking classifier provided accuracy = 98.6% and precision score =0.98 for detecting spam and ham messages accurately under similar setup as a voting classifier. We have also observed that when a balanced training data set was provided, the overall performance of the proposed model improved.



	Text	topic	propensity
0	go jurong point crazi avail bugi n great world la e buffet cine got amor wat	topic_6	0.9558536410331730
1	ok lar joke wif u oni	topic_1	0.8893935084342960
2	free entri wkli comp win fa cup final tkt 21st may text fa 87121 receiv entri question std txt rate c appli 08452810075over18	topic_2	0.5107525587081910
3	u dun say earli hor u c alreadi say	topic_0	0.4436487853527070
4	nah think goe usf live around though	topic_7	0.6538131833076480
5	freemsg hey darl week word back like fun still tb ok xxx std chg send rcv	topic_5	0.4716442823410030
6	even brother like speak treat like aid patent	topic_3	0.3238268494606020
7	per request mell oru minnaminungint nurungu vettam set callertun caller press copi friend callertun	topic_5	0.9489043951034550
8	winner valu network custom select receivea prize reward claim call claim code kl341 valid hour	topic_4	0.3434039652347560
9	mobil 11 month u r entitl updat latest colour mobil camera free call mobil updat co free 08002986030	topic_5	0.7939292788505550
10	gon na home soon want talk stuff anymor tonight k cri enough today	topic_4	0.4949806332588200

Fig. 5. Model Comparison.

B. Topic Modeling

The proposed model employs the LDA model to categorize spam and ham messages into ten topics, which aids in understanding the structure and content of the message. According to result from Fig. 5., topic 0 focuses on messages that provide an incomplete and ambiguous conversation and has a message count of 768. Topic 4 has 641 messages and highlights promotional messages about prizes and offers. Topic 6 with the most messages (831) focuses on daily conversational messages. Thus, topic modeling proved useful in gaining more insight from messages and understanding their structure.

V. Limitations

Since this data was gathered primarily from students for research purposes, the number of spam messages in the dataset was quite limited. Furthermore, the data was restricted to the student body in UK, resulting in a narrower range of messages covered in the data set. Furthermore, SMS messages are confidential source data, and there were few publicly available datasets with for our experiment. As a result, when using data from different regions, the model's performance may suffer.

VI. Future Scope

Future research on this system can be conducted by:

- Creating a spam detection system using transfer learning of Bert models and comparing its performance to our proposed model[3].
- Using neural network-based approaches such as convolutional neural networks and recurrent neural networks to improve performance even further[4,6].
- Implementing the proposed model on multiple combined data sets from different regions to further evaluate model performance.

VII. Ethics

To ensure the SMS spam collection dataset copyrights for academic research projects, the dataset was obtained from Kaggle in accordance with the research's terms and conditions. The data is also made available to the public via the UCI machine learning repository for academic research purposes.

VIII. Conclusion

In this paper, we propose an effective model for classifying spam and ham SMS from the SMS spam collection dataset using natural language pre-processing techniques combined with supervised machine learning algorithms for spam SMS detection. The proposal compared 13 different sentiment classification algorithms on various setups for classifying spam and ham messages that are trained and tested. The results of each classification algorithm are compared to determine the performance of the proposed model. This proposal will assist mobile network providers to find more insights about the focused topic in a message by using topic modeling and determining the best approach for detecting spam SMS to provide data privacy to its subscribers.

References

[1] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Modern Deep Learning Research," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 09, pp. 13693–13696, Apr. 2020, doi: 10.1609/aaai.v34i09.7123.

[2] G. Ubale and S. Gaikwad, "SMS Spam Detection Using TFIDF and Voting Classifier," in 2022 International Mobile and Embedded Technology Conference (MECON), Mar. 2022, pp. 363–366. doi: 10.1109/MECON53876.2022.9752078.

[3] V. S. Tida and S. Hsu, "Universal Spam Detection using Transfer Learning of BERT Model." arXiv, Feb. 07, 2022. doi: 10.48550/arXiv.2202.03480.

[4] S. Annareddy and S. Tammina, "A Comparative Study of Deep Learning Methods for Spam Detection," in 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Dec. 2019, pp. 66–72. doi: 10.1109/I-SMAC47947.2019.9032627.

[5] S. Y. Yerima and A. Bashar, "Semi-supervised novelty detection with one class SVM for SMS spam detection," in 2022 29th International Conference on Systems, Signals and Image Processing (IWSSIP), Jun. 2022, vol. CFP2255E-ART, pp. 1–4. doi: 10.1109/IWSSIP55020.2022.9854496.

[6] S. Gadde, A. Lakshmanarao, and S. Satyanarayana, "SMS Spam Detection using Machine Learning and Deep Learning Techniques," in 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Mar. 2021, vol. 1, pp. 358–362. doi: 10.1109/ICACCS51430.2021.9441783.

[7] S. Bosaeed, I. Katib, and R. Mehmood, "A Fog-Augmented Machine Learning based SMS Spam Detection and Classification System," in 2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC), Apr. 2020, pp. 325–330. doi: 10.1109/FMEC49853.2020.9144833.

[8] H. Jain and R. K. Maurya, "A Review of SMS Spam Detection Using Features Selection," in 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), Jul. 2022, pp. 101–106. doi: 10.1109/CCiCT56684.2022.00030.

[9] "Spam Detection with Topic Modelling." https://kaggle.com/code/nbuhagiar/spam-detection-with-topic-modelling (accessed Dec. 09, 2022).