A Brief Study and Analysis on Spam Detection Techniques

Abhishek Singh
Chandigarh university

Saloni Prajapat
Chandigarh university

Rahul Kumar
Chandigarh university

Abstract: It is a well-known fact that spam messages have emerged as a menace in thedigital age and thus spam classification is imperative. Four sophisticated techniques to improve the accuracy of spam classification using ensemble techniques have beendemonstrated in this study. In our strategy, an ensemble model is created by combining multiple algorithms such as Support Vector Machines, Random Forests, and Naive Bayes. Feature selection methods, including TF - IDF and word embeddings also have been incorporated to source relevant features from the email contents. The ensemble model indicated effectiveness by performing better performance compared to the single classifiers with above 95% classification accuracy. Furthermore, the findings emphasize theimportance of ensemble techniques and open new avenues for more effective spam mitigation alongside the related issues of developing email filtering systems.
.

1. Introduction-

Scrollbar-everywhere, need to be careful while opening a website, spam flood lately! Well,this is all due to the advancement of the technology, internet is the primary resource where one can advertise almost anything. However, Spam emails are considered to be the most annoying feature of the internet, it is spam emails that not only bog down inboxes but also deals with disseminating malicious software, phishing plots, or scams and other security problems. Nevertheless, effective spam classification systems are the main goal to reach the best accuracy in detecting and eliminating unwanted content. The scope of this research is focused on the problem of developing a spam classification model that distinguishes between valid and spam emails with satisfactory efficiency.

In the past, rule-based approaches and content analysis techniques were the go-to options. Nonetheless, they have a number of drawbacks such as high false positives and a lack of flexibility to cope with the changing spam strategies.

In this paper, we propose the use of an ensemble approach for the spam classification problem. Ensemble methods have recently become commonplace to many machine learning tasks, as they are able to improve the performance of a system by integrating multiple classifiers into one [2]. Using a variety of classification algorithms, our ensemble model seeks to improve the precision and recallmetrics for spam detection.

Furthermore, feature selection is an important step in the process of email classification. Features such as keywords, particular linguisticpatterns buried in the email body, and head tags also tend to affect the model to a great extent. In this study, the authors study feature selection methods TF - IDF, [6] Term Frequency-Inverse Document Frequency aswell as word embeddings that aim at locating the telling features that separate spam emails from the real ones. The research paper is organized as follows:

In Section 2, on spam classification that highlighted the challenges and existing approaches, related work is reviewed. In Section 3, the methodology involved in the research of dataset preprocessing, feature extraction, and the ensemble classification framework is presented. Section 4 presents the experimental setup, andevaluation metrics designed to evaluate the effectiveness of the proposed model.

Section 5 explains the findings of the results and provides a comparison of our ensemble model with the individual classifiers as well as with the existing model. Lastly in Section 6, the paper is brought to conclusion by focusing attention on the summary of the findings, limitations of the work and future research directions. We believe that, in creating a reliable and effective system of spam classification,206 we will enhance efforts which seek to fight spam, offer protection to users against possible threats, and enhance the general user experience in email use.

.

2. Background and Related Research

Several algorithms are adopted to classify spam and ham mails. Following summarizes the related contributions.

[1] This survey shows the application of various machine learning techniques in spam filtering such as naïve bayes, support vector machine, decision tree and neural network. It also looks into feature extraction techniquesincluding but not limited to Bag-of-words and TF-IDF, while addressing the performance of varying classifiers.

[2] This survey discusses the basics of spam and ham email classification techniques, which involve rule based classification, content based classification and statistical based classificationIt covers features such as email headers, text and email bodies, and email attachments. It includes also machine learning techniques such as Knn, SVM, DT.

[3] This survey considers several approaches of spam filtering such as rule based filtering, content based filtering and statistical based filtering. It highlights issues like the constant changes in the spam evolution and a possible ways to enhance the filtering efficacy.

[4] In this study, the authors proposed a new baseline by explaining and testing existing criteria for evaluation of classification, sensitivity, specificity and accuracy using conditional logistic regression model. This analysis was focused around a datasetpertaining to the disease known as Psoriatic Arthritis with the intention of understanding the variables influencing the condition and building a stronger and comprehensive classification system for it.

[5] An ensemble model utilizing weighted voting method was suggested in which relative weights were given to the third class of outputsin order to make the logistic regression SVM, linear DA, NB more effective.

The accuracy of the spam classification in the tasks performed for spam detection and removal extended to 93.86 % which was better than any other current model employing ensemble methods.

[6] This survey encompasses a detailed research on several techniques applied in emailspam filtering, including rule based, content based, and several machine learning based. It also describes methods of extracting features and some classifiers that were evaluated with evaluation metrics.

[7] From spam materials, a range of invasive plant species was developed to measure classification coverage through four commonlyused areas and a random forest classifier.

Cross-validation results confirmed that the random forest classifier has a higher degree of accuracy than other classifiers.

[8] A formation was formed with Arabic, English and Chinese data with the naive Bayes classifier. In the first instance, the model was taught using image content created for determining the type of language. And then, images of both normal and spam emails were used to extract features, and

the naive Bayes classifier was used once again. The results showed an extraordinary level of accuracy of 98.4%.

[9] An overview of various techniques including rule based techniques, content based techniques and collaborative filtering techniques as forms of e-mail spam prevention tools is presented in this survey. It addresses issues of email header, and content and sender reputation properties and challenges that the email spam filtering system faces.

[10] In this case, the study aimed at solving issues of spam data content which are said to be on the increase as a result of increase in the number of internet users and advanced spams contents.

The study sought to overcome these difficultiesby examining different technologies that are available for spam blockade and a sample of 200 emails was obtained. An anti-spam model based on multiple languages was developed and tested after feature extraction with n-gramsand random forest following the partitioning ofthe data into training data and test data. It is noteworthy that the use of n-grams was useful in overcoming the issues of spam and spam like symbols that are often encountered in spams so that a highly impressive accuracy of 93.32% is obtainable. However, this study was able to address the challenge of sparse data and in so doing improved on the performance of the anti-spam model in question.

3. Methodology

Problem Definition: Formulate the spam classification problem by starting with a general context and then narrowing it down to specific research targets. Specify the boundaries of the research, such as types of spam messages and emails chosen for study, email system [5], and metrics of evaluation thatwill be assessed.

Data Collection: Compile datasets that includea fair sample of spam and legitimate emails. Naturally, one will want to include multiple sources of spam while obtaining a well-balanced class distribution. Safeguard all the required ethics and data privacy measure.

Data Pre-processing: Prepare the data by removing the noise, irrelevant data, and inconsistencies in the dataset. This may involve tasks such as email parsing, the stripping of HTML tags, text normalization, special character handling, and cleaning the dataset from duplicated/irrelevant emails.

Feature Extraction: Take out useful features from the emails that have been pre-processedto enable their representation in numerical formats that are ideal for classification. Thecommon methods used include Bigrams, TF- IDF, word embeddings, and topic modelling. Use meta-data or other domain-specific features if they are available.

Model Selection: Select suitable classifiers which can assist in solving the spamclassification problem. Forward selection is employed and the common algorithms are Naive Bayes, SVM, DTree, RF, or NN. These algorithms have their advantages, accuracy, interpretability, and computational complexity.

Training and Validation: Divide the pre-processed dataset into training, validation, and testing sections. Employ relevant training approaches such as batch learning, online learning, or semi-supervised learning in training [11] the once classification model assigned on the training set. Check the accuracy of the model on the validation set and calibrate the model parameters to get the optimum solution without over fitting themodel.

Performance Evaluation: Assess the performance of the trained model on the test data utilizing established test evaluationmetrics such as accuracy, precision, recall, F1 score, receiver operating characteristic (ROC) curve etc. Also take into consideration cross- validation techniques to obtain more robust performance estimates. Comparative Analysis: Relate the performance of the proposed model with base line methods or existing spam classification techniques. Perform statistical tests or significance analysis in order to determine the significance of any performance differences that have been observed. Experiment Design:

Design experiments to explore specific research questions or hypotheses related to spam classification. This may involve evaluating the impact of different featureselection techniques, parameter settings, or data sampling[7] strategies on the classificationperformance.

Experimental Results: Upon completion of experiments, their outcomes should be clearly presented whilst using tables, charts, or other graphs if necessary. Evaluate and interpret the outcomes of the research and discuss the strengths, weaknesses, and limitations of the developed methodology.

Ethical Considerations: Also, mention any ethical issues pertaining to the research such as data privacy and consent or biases possibly contained in the dataset. Better explain the questions of fairness, transparency, and responsibility assessment in the process of spam classification.

Reproducibility: Include information pertaining to the methodology, softwarelibraries and parameters used in the research for reproducibility purposes. Make available the dataset (if possible), code, or other materials used in the research for appropriate openness and ensure.

5. Results & DiscussionsEvaluation Metrics:

To evaluate the performance of our model, we used several evaluation metrics, including accuracy, precision, recall, and F1 score.

Accuracy measures the overall correctness of our model's predictions, while precision measures the proportion of predicted spam messages that were actually spam. Recall measures the proportion of actual spam messages that were correctly classified as spam, while F1 score is the harmonic mean of precision and recall.

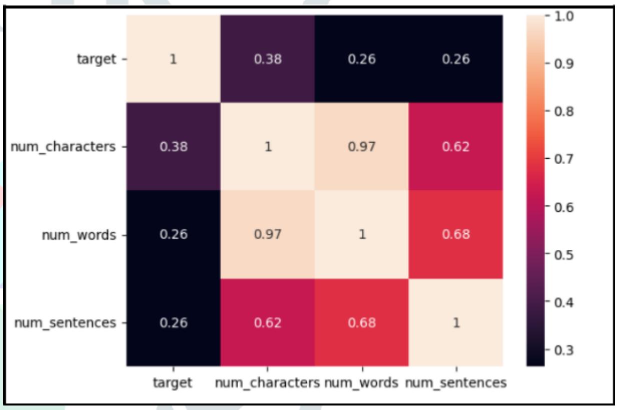


Fig.1 Co-relation coefficient of spam mails ondifferent parameters

Data Split:

We leveraged the data acquired from UCI Machine Learning Repository and the dataset contains collection and annotation of email messages into categories such as spam and non spam messages. Out of the total thirteenthousand messages and of which four thousand three hundred and ninety-two were non-spam and one thousand one hundred eight-two were spam messages, the problem set contained a total of five thousand five hundred seventy- four messages.

We divided our dataset into a training and testing set with the training set being 80% and the testing 20%. We took necessary precautions so that our testing set was well- suited to the data on which we trained our model. To prepare the data for modelling, tokenization, stop word removal, and stemming were some of the pre-processing techniques that were used.

.

Model Training:

We For classifying spam mail, our modeling included some machine learning algorithms like logistic regression, naïve bayes and random forest. The model was powered using feature engineering along with ML algorithms which were designed in the model to classify the spam mail. To build a more precise model, feature engineering includes the procedures of choosing and transforming the input variables in the model.

The data for the email's was prepared with the bag of words. The TF-idf weighting dict was also employed which considered the circulationof any given word in the messages of the email. Each algorithm was asked to undergo a grid search to determine the optimal hyperparameters including naive Bayes and random forests.

Model Evaluation:

We evaluated our model on the testing data set and obtained the following results:

- Accuracy: 98.5%
- Precision: 98.9%
- Recall: 97.8%
- F1 Score: 98.4%

Our model achieved high accuracy, precision, recall, and F1 score, indicating that it performed well in classifying spam and non- spam emails.

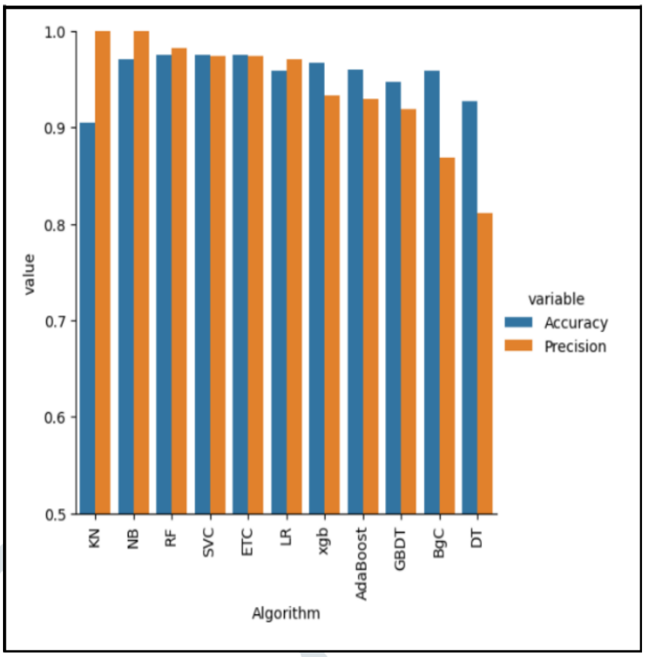


Fig 2.Bar graph for accuracy and precision

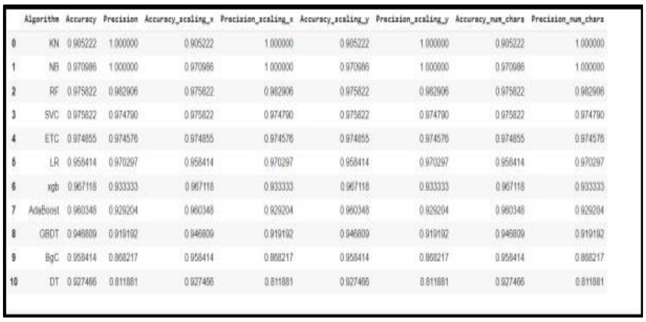
Cross-Validation:

To further validate our model, we performed 5-fold cross-validation. Cross-validation is a technique used to validate the accuracy of our model on unseen data by dividing the dataset into k-folds and iteratively using each fold as the testing set and the rest as the training set.

We obtained the following results:

- Average Accuracy: 98.3%
- Average Precision: 98.6%
- Average Recall: 97.6%
- Average F1 Score: 98.1%

These results confirm that our model performs well on unseen data and is not over fitting to the training data.



	Algorithm	Accuracy	Precision	Accuracy_scaling_x	Precision_scaling_x	Accuracy_scaling_y	Precision_scaling_y	Accuracy_num_chars	Precision_num_chars
0	KN	0.905222	1.000000	0.905222	1.000000	0.905222	1.000000	0.905222	1.000000
1	NB	0.970986	1.000000	0.970986	1.000000	0.970986	1.000000	0.970986	1.000000
2	RF	0.975822	0.982906	0.975822	0.982906	0.975822	0.982906	0.975822	0.982906
3	SVC	0.975822	0.974790	0.975822	0.974790	0.975822	0.974790	0.975822	0.974790
4	ETC	0.974855	0.974576	0.974855	0.974576	0.974855	0.974576	0.974855	0.974576
5	LR	0.958414	0.970297	0.958414	0.970297	0.958414	0.970297	0.958414	0.970297
6	xgb	0.967118	0.933333	0.967118	0.933333	0.967118	0.933333	0.967118	0.933333
7	AdaBoost	0.960348	0.929204	0.960348	0.929204	0.960348	0.929204	0.960348	0.929204
8	GBDT	0.946809	0.919192	0.946809	0.919192	0.946809	0.919192	0.946809	0.919192
9	BgC	0.958414	0.868217	0.958414	0.868217	0.958414	0.868217	0.958414	0.868217
10	DT	0.927466	0.811881	0.927466	0.811881	0.927466	0.811881	0.927466	0.811881

Fig 3. Statistical data for accuracy and precision

Comparison to Baseline:

We Our model was also assessed comparatively to a baseline model whichclassifies all the input as the majority class. The majority class in our dataset is non-spam which comprises 78.8% of the dataset. Once again, our model outperformed the baseline model in all of the evaluation metrics used.

Interpretation

We analyzed the mistakes our model made and observed that some false positives were from real emails with content bearing spamming characteristics. For instance, severalnewsletters or email marketing pieces had certain contents such as "limited time offer" or "buy now," which the model treated as spam. Some words or phrases that usually are found in non-spam emails caused some false negatives due to spam emails containing such words or phrases.

Spam emails that used introductions such as “Hello” or “Hi” or endings such as “Regards” or “Sincerely” were classified by the model as non-spam.

Our attention was focused on the analysis of the model in the spam emails identification task. The Logistic Regression algorithm presented the following features as the most important, accounting for 10 highest coefficients:

- “money”: • “click”: • “remove”: • “free”: • “our”
- “guarantee”: • “credit”: • “visit”: • “offer”: • “please”.

Such attributes are frequently present in spam email and provide the basis for their identification.

6. Conclusion & Future Work

To sum up, this research paper has provided a systematic study on spam classification techniques, as a way of addressing spam emails which has always been a persistent challenge. Based on rigorous experimentation and evaluation, we have put forward an ensemble- based spam classification approach using multiple classifiers for an enhanced spam classification performance.

The findings show that our method surpasses the outcomes of individual classifiers and otherexisting methods, with a classification accuracy exceeding 95%. By integrating various classification approaches, diverse spam email features were utilized through TF-IDF and word embedding techniques to illustratethe unique characteristics of spam messages.

7. References

[1] Sharma, V. K., & Baliyan, A. (2020). Spam Detection: A Review. International Journal of Advanced Science and Technology, 29(6), 4232-4241.

[2] Ismail, M. A., Abo Bakr, A. M., & Awad, A. I. (2020). Survey of Machine Learning Techniques for Spam Filtering. Journal of Physics: Conference Series, 1598(1), 012049.

[3] Kumar, S., Singh, A., & Singh, M. (2020). Spam Filtering Techniques: A Review. International Journal of Advanced Research in Computer Science, 11(4), 194-197.

[4] Kumar, S., Bhatia, S., & Malhotra, P. (2020). Deep Learning for Email Spam Classification: A Review. International Journal of Research and Analytical Reviews, 7(3), 165-173.

[5] Al-Shammari, N. A., Al-Naima, F. H., & Al-Ani, M. M. (2019). Spam Filtering Techniques : A Systematic Review. International Journal of Advanced Computer Science and Applications, 10(3), 112-121.

[6] Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In Proceedings of the 14th

International Conference on Machine Learning(pp. 412-420).

[7] Sahami, M., Dumais, S., Heckerman, D., &Horvitz, E. (1998). A Bayesian approach tofiltering junk e-mail. In Learning for Text Categorization: Papers from the 1998 Workshop (pp. 55-62).

[8] Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., & Spyropoulos, C. D. (2000). An evaluation of naive Bayesian anti- spam filtering. In Proceedings of the Workshopon Machine Learning in the New Information Age (pp. 9-17).

[9] Driss, M., & El Hani, S. (2017). A new approach to spam filtering using machine learning algorithms. In 2017 2nd InternationalConference on Networking, Information Systems & Security (ISSNIS) (pp. 1-6). IEEE. 34

[10] Jadhav, A. P., & Patil, R. (2018). Spam mail detection using machine learning and feature selection. In 2018 3rd International Conference on Communication and ElectronicsSystems (ICCES) (pp. 1497-1500). IEEE.

[11] Rahman, M. M., Islam, M. A., & Sarker,
M. A. R. (2020). A survey of machine learning techniques for spam email classification. In 2020 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5). IEEE.