Content-Based Anti-Spam Detection Using ML and NLP

Prayag Gavshinde[1] Raghav Agrawal[2] Rishi Somani[3] Sambhav Jain[4] Dr. Asif Ali[5]
[5]Associate Professor
[1,2,3,4,5]Department of Information Technology
[1,2,3,4]Acropolis Institute of Technology and Research, RGPV, Indore, India [5]Acropolis Institute of Technology and Research Indore, India

Abstract— Spam filtering is a widely discussed and studied topic In the field of pattern classification. Emails, SMS as well social media comments can be filtered as spam and not spam based on many features like frequency or occurrence of few words than content, the length of the e-mail, SMS, or the domain from which it is being sent. Based on these basic characteristics, researchers have come up with many techniques to identify content as spam and non-spam content. In this project, we aim to implement and evaluate three major e-mail spam filtering algorithms. They are, Naïve Bayes method, k-Nearest Neighbors, and Support Vector Machines.
Keywords: Text Classification, NLP, Machine Learning, Deep Learning

I. INTRODUCTION

The internet, as we know it, is a widely used platform for sharing knowledge and resources. Going five decades back in time, one couldn't even have imagined that sharing information, be it texts, images, ideas or even seeing one another virtually would be such an easy task, as it is now. However, with this ability to exchange ideas, thoughts, messages, and news at lightning speed, comes the threat of falling prey to malicious intentions of people who use the World Wide Web for wrong purposes such as fraud, cyber-bullying, and other forms of cyber-crime. E-mails and SMS are a major way of communicating over the internet. Currently, E-mail is one of the most important methods of communication. However, the increase in spam emails causes traffic congestion, decreasing productivity, and phishing, which has become a serious problem for our society. And the number of spam emails is increasing every-year. Therefore, spam email filtering is an important, meaningful, and challenging topic. The aim of this research is to find an effective solution to filter possible spam emails. And as we know, in recent days, there are many techniques that spammers use to avoid spam-detection such as obfuscation techniques and many more.

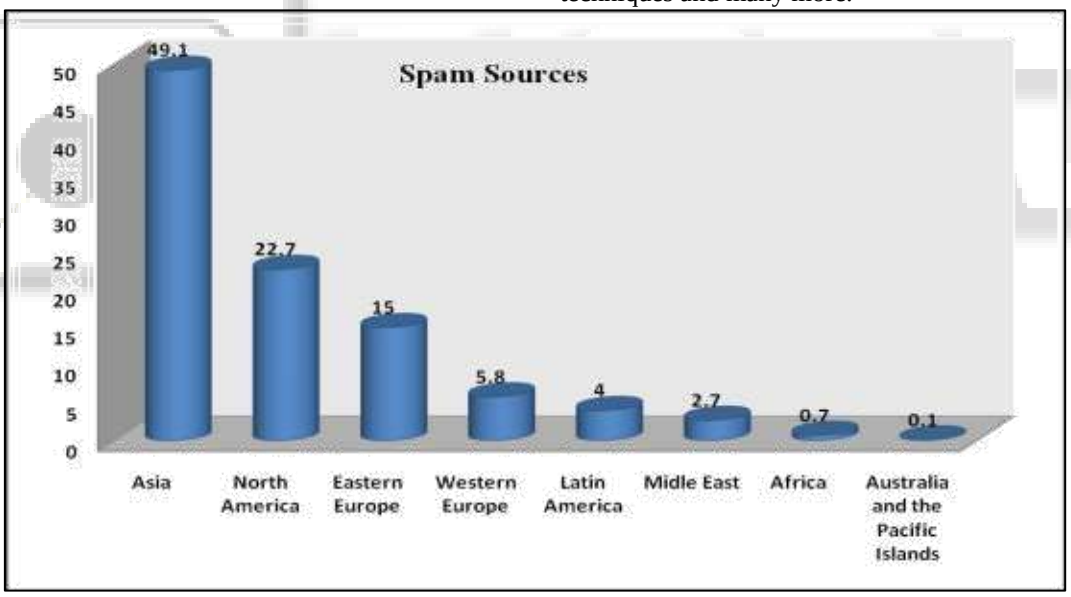

Fig. 1: Comparing Spam generation across various continents

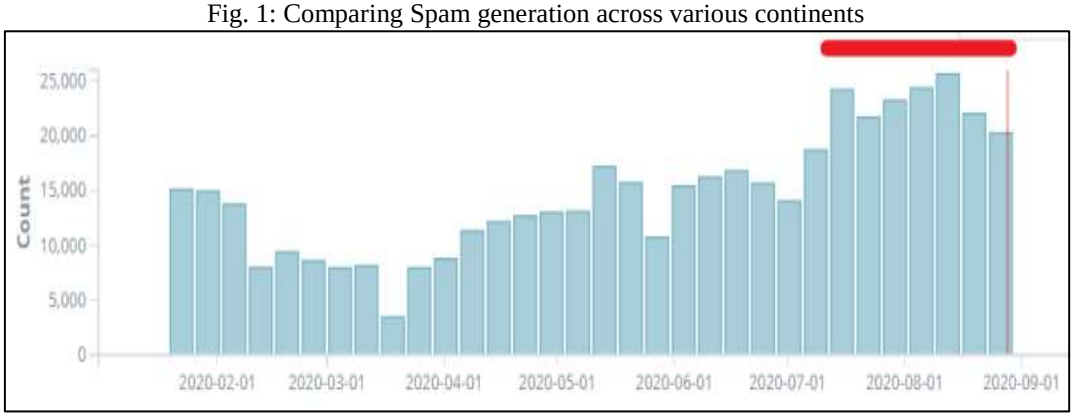

Fig. 2: Spam content generation across India in last year.

Asia is the largest continent in the generation of Spam as compared to others. In the last 2-3 decades the data which is generated is twice everyday and near about 40 to 50 percent data at data centers is spam and fake which is used by hackers and many others to cheat people in some way.

II. EXISTING SYSTEM

The current systems of spam detection are solely dependent on three main methods:-

A. Linguistic Based Methods

Humans can comprehend linguistic constructs and their interpretations, but machines can't, and so machines are taught some language in order to help them comprehend linguistic constructs. These techniques are used in search engines to determine the next term in an unfinished sentence. They are split into two Unigrams (Words one by one) and two Bigrams (Words two at a time). As every term has to be remembered, this approach is not as reliable and time-intensive.[7]

B. Behavior-Based Methods

It is based on Metadata. This method requires users to create a set of laws, and users need to have extensive knowledge of such laws. It needs reformulation because the characteristics of spam shift over time and the laws need to be modified accordingly. As a consequence, it is mostly user-dependent, and still human needs to examine more details. [3]

C. Graph-Based Methods

In this approach, by integrating many, heterogeneous details into a single graphical representation, unusual patterns are detected in the data that shows spammer behaviors by running graph-based anomaly detection algorithms for graphical representation. This approach is not reliable, so it is challenging to detect false opinions.[8]

Existing System working on above methods for Spam classification:-

No.	Existing Software/ System	Description	Limitation
1.	*Spam signature generation based framework [4]*	Aim to analyze behaviour of spam in a network through characteristics and properties works on AutoRE framework.	Do not provide any clarity how well it is performed in real time for spam campaigns.
2.	*Characterizing botnets from Email spam records*	The technique uses traces of spam email to map botnets in groups.	It lacks to provide online detection and monitoring of the networks.
3.	*Botsniffer*	Prototype system used to capture spatial temporal correlation in network traffic and utilize statistical algorithms.	It requires observing multiple rounds of response crowds and in less crowd accuracy suffers.
4.	*Botminer*	This works under the influence of boot master and utilizes all resources to counter denial of services (DOS), spam attacks.	It tends to promote previous patterns.

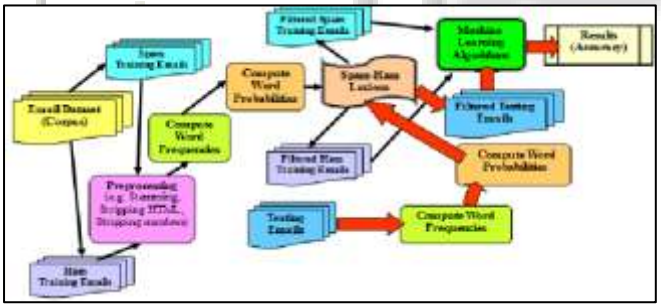

Fig. 3: Block Diagram for text classification

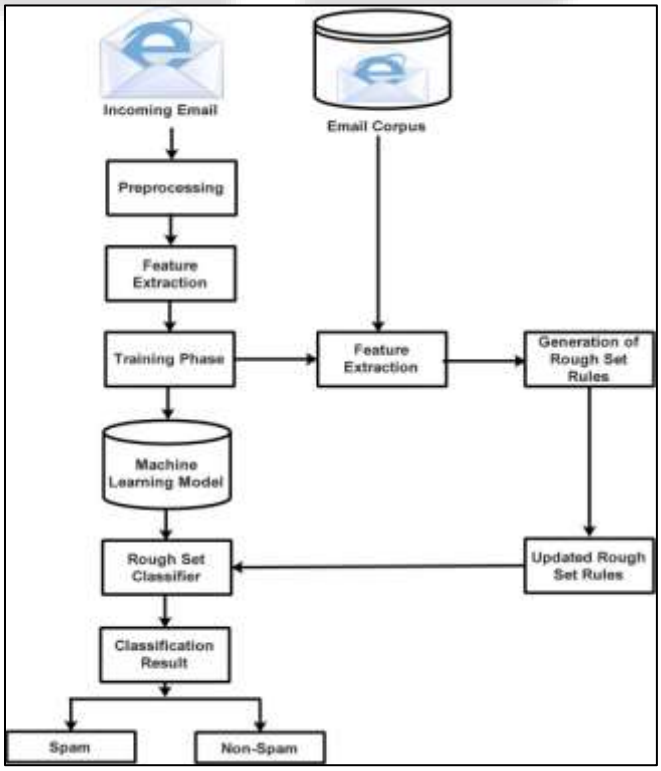

Fig. 4: Process Flow chart to classify spam

III. PROPOSED SYSTEM

A. Requirements:

1) Software- Python IDE and its libraries, Jupyter Notebook, Web-Framework (Flask), cloud Server (Heroku)
2) Hardware-
   - Modern Operating System:
     - Windows 7 or 10
     - Mac OS X 10.11 or higher, 64-bit
     - Linux: RHEL 6/7, 64-bit (almost all libraries also work in Ubuntu)
   - 4 GB RAM
   - x86 64-bit CPU (Intel / AMD architecture)

The proposed framework consists of a set of modules that are implemented:

B. Dataset Extraction

First data is collected from the dataset, in our case which is email, SMS, and comments. After collecting the data, it is cleansed by getting rid of extra spaces, removing duplicates, and many more.

C. Generalize Data

All cleaned data collected and generalized regardless of whether they are spam or not based on different parameters and a corpus is created by extracting the features. By generalizing the data a lot of time can be saved.

D. Implementing ML Algorithms

The ML algorithms are implemented in this stage by segregating the content into spam content and original content. ML algorithms including Decision Tree, Naïve Bayes, K-nearest neighbor, and support vector machine are used.[6]

E. Generating Spam Text Data and information about the Content

After the ML algorithms have been implemented the spam content is identified and obtained, and the information about the content to provide the user the confidence to believe or not to believe. With the help of this information, all the content can be analyzed and gain confidence.

The system that is proposed in this paper combines Machine Learning algorithms which is a supervised classification algorithm with NLP concepts to categorize and detect spam among all existing SMS, emails, and comments on the youtube dataset. There are major features used in the algorithm which includes many NLP concepts like stop-words, lemmatization.

The proposed framework is introduced with the aid of two key applications, one is anaconda prompt which is exactly similar to the usual command prompt and the other is Jupyter, an integrated python development environment. The anaconda prompt is used for running anaconda and conda commands without changing the directories and to access the localhost by connecting the file folder to it and downloading and extracting packages to implement the framework. [10]

IV. RESULTS AND DISCUSSION

The results are based on the performance of Machine Learning Algorithms as various Performance Metrics which include accuracy score, Precision, Recall, F-beta score for an imbalanced dataset of SMS.

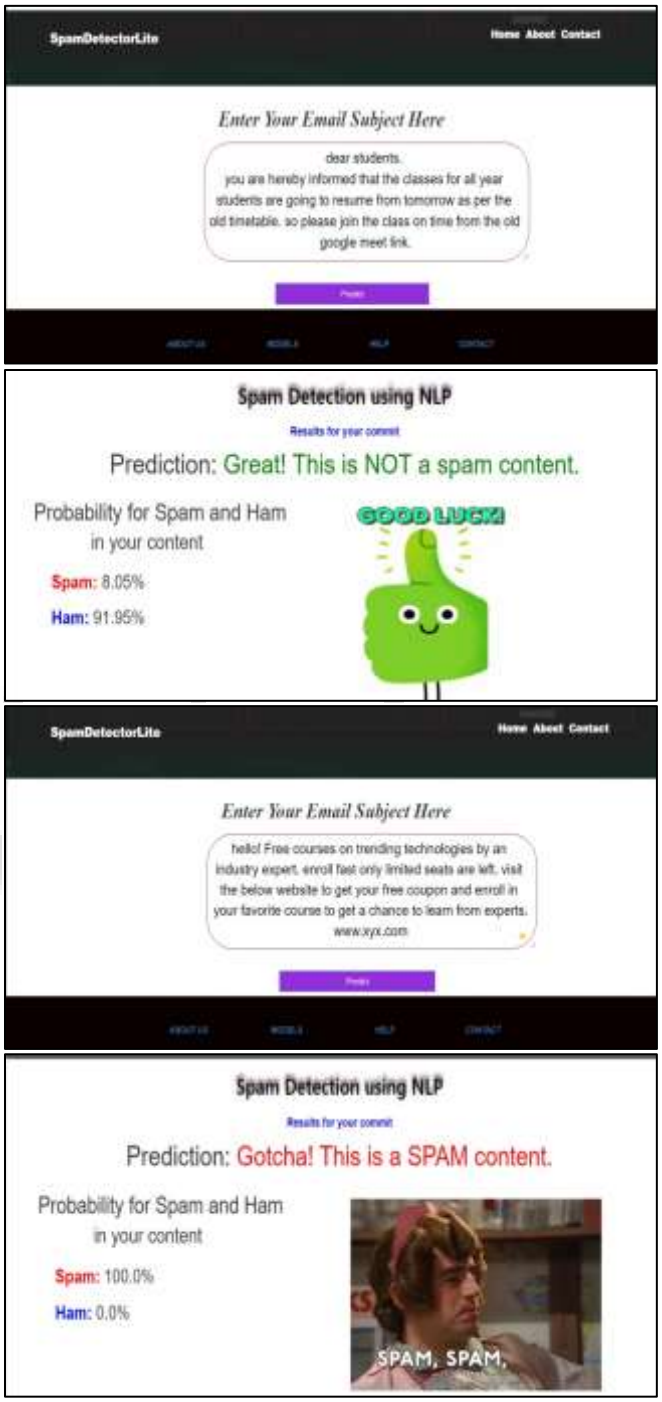

Fig. 5: Classifying content as spam and Ham

1) Test Case-1) Email Spam Classification

Models Used	Accuracy	Precision	Recall
Gaussian Naive Bayes [1]	95.41%	1 - 0.97 0 - 0.89	1 - 0.97 0 - 0.92
SVM (support vector machine)	95.96%	1 - 0.97 0 - 0.92	1 - 0.98 0 - 0.90
Decision Tree	96.43	1 - 0.96 0 - 0.93	1 - 0.97 0 - 0.91

Naive Bayes (MultinomialNB)	98.24%	1 - 0.94 0 - 1.00	1 - 0.99 0 - 0.98

2) Test Case-2) SMS Spam Classification

Models Used	Accuracy	f beta-score
Logistic Regression	97.31%	96.23
SVM(support vector machine)	97.48%	96.42
Decision Tree	96.58%	95.87
Naive Bayes(MultinomialNB)	97.86%	96.14

3) Test Case-3) Youtube Comments Spam Classification

Models Used	Accuracy	Precision	Recall
Recurrent Neural network	94.22%	1 - 0.942 0 - 0.93	1 - 0.941 0 - 0.947
SVM(support vector machine)	93.71%	1 - 0.92 0 - 0.93	1 - 0.913 0 - 0.926
Decision Tree	87.43%	1 - 0.908 0 - 0.91	1 - 0.937 0 - 0.91
Naive Bayes (MultinomialNB)	89.24%	1 - 0.91 0 - 0.913	1 - 0.925 0 - 0.93

The final Algorithm Used for Email and SMS is the Naive Bayes theorem based on its performance and other metrics as well. Support vector is also performing very well on its edge but during the deployment process, there are some errors in parameter tuning of the Support Vector Machine. And in classifying youtube comments bi-directional LSTM is used RNN. The overall average accuracy of all three models is approximately 95 percent which is good for classifying the content.

B. Advantages

- Feature Engineering is available. Therefore, the features of raw data can be easily extracted with the help of data mining. It is used to improve the performance of Machine learning algorithms.
- Each and every data obtained is accurate.
- Spam Features are a built-in function.
- It is a statistics-based approach that Supports review-centric spam detection as well as Supports reviewer-centric spam detection.

V. Conclusion and Future Work:

In this paper, we identified the spam and phishing content present in a data user receives or sends with the help of machine learning algorithms and NLP concepts. By reviewing the spam, the entire details about the content classification as spam and not spam as well the probability of the content to be spam are accessed and displayed, which in turn helps in determining other spam, spammers and their way of writing messages.

We considered two attribute sets which include content and extracted spam corpus. The content is determined with the help of average content similitude, maximum content similitude, the ratio of exclamation sentences, and the ratio of first personal pronouns. The probability is determined according to the content similitude to corpus with spam and ham words on which algorithm is pre-trained. Thus, making it a very effective and accurate spam detection framework.

References

[1] Nurul Fitriah Rusland, Norfaradilla Wahid, Shahreen Kasim, Hanayanti Hafit, "Analysis of Naive Bayes Algorithm for Email Spam Filtering across Multiple Datasets"
[2] J. Rout, S. Singh, S. Jena, and S. Bakshi, "Deceptive Review Detection Using Labeled and Unlabeled Data".
[3] T.S. Guzella, W.M. Caminhas A review of machine learning approaches to spam filtering Expert Syst. Appl., 36 (7) (2009), pp. 10206-10222 Article
[4] N. Pérez-Díaz, D. Ruano-Ordás, F. Fdez-Riverola, J.R. Méndez Rough sets for spam filtering: Selecting appropriate decision rules for boundary classification Appl. Soft Comput., 13 (8) (2012), pp. 1-8 View Record in Scopus
[5] J. Han, M. Kamber, J. Pei Data Mining: Concepts and Techniques. Elsevier (2011) Google Scholar
[6] Shrawan Kumar Trivedi, "A Study of Machine Learning Classifiers for Spam Detection".
[7] G.V. Cormack Email spam filtering: a systematic review Found. Trends Inf. Retr., 1 (4) (2008), pp. 335-455 CrossRefView Record in Scopus
[8] A. Bhowmick, S.M. Hazarika Machine Learning for E-Mail Spam Filtering: Review, Techniques and Trends arXiv:1606.01042v1 [cs.LG] 3 Jun 2016 (2016), pp. 1-27 CrossRef
[9] I. Katakis, G. Tsoumakas, I. Vlahavas Email mining: emerging techniques for email management A. Vakali, G. Pallis (Eds.), Web Data Management Practices: Emerging Techniques and Technologies, Idea Group Publishing, USA (2007) chap 10 Google Scholar
[10] T.A. Almeida, A. Yamakami - Content-based spam filtering: The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona. (2010), pp. 1-7. View Record in ScopusGoogle Scholar

