Multi-Stage Spam Detection in E-mails via Machine Learning

Dr. Subedha V, Deebika, Moulika A K, Ms.Tamilselvi K

Dr. Subedha V, HOD/CSE, Anna University, Panimalar Institute of Technology
Deebika S, Anna University, Panimalar Institute of Technology
Moulika A K, Anna University, Panimalar Institute of Technology
Ms.Tamilselvi K,AP/CSE, Anna University, Panimalar Institute of Technology

Abstract:

Spam mails or junk mails are usually employed for the purpose of marketing or advertising a product or an organization. Accumulation of spam mails result in the wastage of in-built memory in email applications. Spam mails can be both harmful and harmless. Harmful spam mails might contain malicious links, which on accessing might steal user data whereas harmless spam mails are commonly employed for promotional purposes. Everyday, spam mails are shared in bulk to several email users, irrespective of the content and the language. Considering the problem, this paper has come with the approach of multi-stage spam mail detection methodology that can beapplied to a wide range of spam mails and is implemented using Naïve Bayes Classifier Algorithm.

The multi-stage spam detection system uses different datasets that store spam keywords belonging todifferent languages respectively. Every set of mails pass through four stages of detection, through whichthe junk mails are filtered out.

Keywords-Spam mails, spam detection, junk mails, multistage detection, machine learning

1.INTRODUCTION

Electronicmail,alsoknownasemail,isamedium of communication between two or more people over the Internet. If any mail is received from a user, who is not available in our close circle, then the message is a spam. Notallthespam mails are malicious. Some mails just promote or advertise any product, institution, or organization. An increase in the count of spam mails could drain the storage space.The othercategory ofjunk mails includesmasturbatingcontents,virus infected files, URLsthatstealuserdata,whichare considered illegalandresultsincybercrime [1].

Agraphicalreport of the recent survey conducted by the organization *Statista* is displayed below.The graph displays the top 10 countries where individual email users receivethemaximumnumberofspammailsevery day. Summarizing the survey, The United States of America stands first with an approximate count of eight billion.Theabove survey inferences that Indian email users are receiving comparatively lesser number of junk mails.

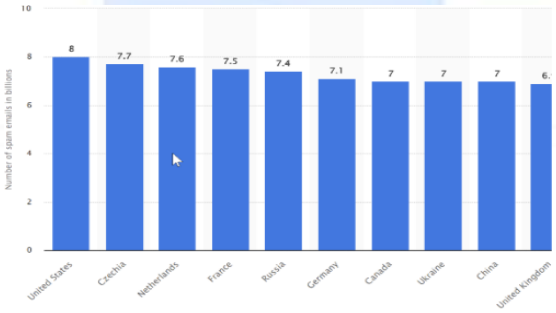

Fig 1: Survey report by Statista

2. CLASSIFICATION

In multistage spam detection, Naïve Bayes algorithm is used. Naïve Bayes classifierscategorize mails into spam and ham [14]. Any mail which is sent by the people (or organization) comes under ham. Mails which are seemed to be suspicious and are received from an anonymous
sender is isolated as spam.

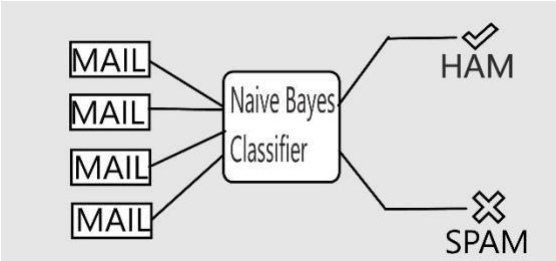

Fig 2: Working of Naïve Bayes Classifier

Spam mails are composed of various categories, such as fake news, advertisements, lottery messages and so on. Some of the spams are defined below:

2.1. FAKE NEWS

A fake news is a false information that is spread by an individual or a group of individuals with an
intention of misleading the public for certain reasons or sometimes for creating nuisance, which is considered a crime.

2.2. LOTTERY MESSAGES

Users receive mails containing an embedded URL with a message, that conveys that the user has won a lumpsum amount in lottery and claim the amount by clicking the link below. These embedded URLs are malicious, such that they tend to steal that person's bank details and other crucial details. There is a possibility of the hacked information getting misused by the cybercriminal.

2.3. ADVERTISEMENT

Advertisement mails are mainly sent for promoting any product, organization and so on. Such mails are harmless and legal.

3. EXISTING SYSTEM

The existing spam email detection system uses datasets that contain keywords that are usually employed in creating spam contents. But the disadvantage is, the datasets used are restricted to a particular language and till now, most of the email applications does not support multilingual spam detection.

4. PROPOSED SYSTEM

Our proposed system passes through four different stages in order to achieve high accuracy in filtering spam mails. The first stage is the basic stage which filters the mails from handles which are marked as spam by the user.
Second stage deals with detection of spam mails belonging to English language, for which the Kaggle dataset is used.
Third stage deals with multilingual spam detection, which uses different language datasets [3]. The final stage includes image-based classification of spam mails. It also includes mails received with suspicious URL embedded in it. Not all the mails with images get stored in spam folder. Certainimages have a high probability of getting flagged as spam. Only such mails will be eliminated by the spam filter.

5.PROCESS FLOW

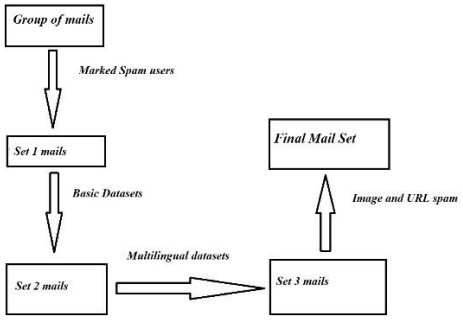

Fig 3: Process flow of the system

6.NAIVE BAYES CLASSIFIER

Naïve Bayes classifier is not a single algorithm, but a group of algorithms based on "Bayes" theorem. Naïve Bayes classifier is primarily employed in text classification tasks such as spam classification. Using Naïve Bayes Classifier, we can classify mails as spam and ham.

Reference	*Literature Survey*	
	Advantages	*Drawbacks*
[1]	*1. Analyses different machine learning techniques and email features in different machine learning approaches.* *2. Analyses about the existing efficient machine learning algorithms for email spam detection.*	*No proper solution is proposed for developing the email spam detecting system.*
[3]	*Easy to detect spam mail received in languages other than English.*	*Unless a sender transmits its first message to a grey list user twice from the same e-mail address within the prescribed time, the message will be rejected(annoying delays).*
[14]	*The delivery approval mechanism is used for evaluating the functions of email agent, thus providing security and confidentiality of documents.*	*The delivery approval mechanism is used for evaluating the functions of email agent, thus providing security and confidentiality of documents.*

6.1. DEFINITION

The machine learning approach used in our project is the Naïve Bayes classifier. "Naive" is the word that expresses the fact that the sample spaces or events that are happening in the specified environment does not depend on one another.

Naïve Bayes classifier is not a single algorithm but a set of algorithms that follow Bayes Theorem. Bayes theorem is used to solve the problems based on conditional probability.

Conditional Probability: Consider two different events A and B. Conditional probability says that there is a possibility that the event A can occur with respect to event B and vice versa, that is one condition can occur with respect to another condition in the same environment.

6.2. GRAPHICAL REPRESENTATION

The given graph represents the number of spam mails filtered with respect the mentioned phases. In each phase, a significant number spam mails are detected and filtered out.

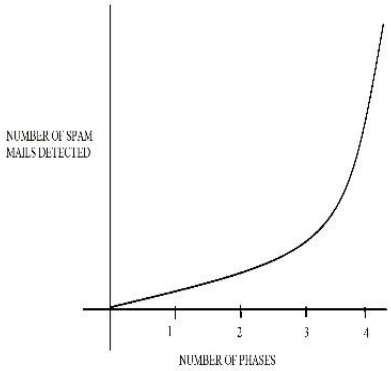

Fig 4: Spam mails detected vs Number of Phases

The filtered-out mails will be increasing exponentially, since the exact number of spam mails received as well as the number of mails eliminated are uncertain. After each phase of filtration, a reduction can be witnessed in the total number of mails received.
The added advantages of this system are:
1. Application of Naïve Bayes Classifier which helps in the efficient detection of junk mails.
2. As the received mails are passing through different phases and each phase expels an appreciable number of spam mails, maximum count of non-spam mails is extracted, thus providing a high accuracy in the elimination of spam mails.

7.CONCLUSION

Spam mails are defined as mails received from unauthorized users which can either be legal as advertising products or illegal as sensitive content sharing, hacking user data via URL. The Multi-Stage spam detection system phases through fourdifferent phases, in which, each phase eliminates a significant number of spam mails are filtered out, irrespective of the language used and content shared. The classification is done using the set of keywords in the database, that are flagged as junk. Although the cost of implementation is high, this system can be used by organizations that aim for achieving high accuracy in vanquishing spam mails. The future researches should be based on improvising this system with respect to performance and cost.

8.REFERENCES

[1] M. RAZA, N. D. Jayasinghe and M. M. A. Muslam, "A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms," 2021 International Conference on Information Networking (ICOIN), Jeju Island, Korea (South), 2021, pp. 327-332, doi: 10.1109/ICOIN50884.2021.9334020.

[2] K. Agarwal and T. Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 685-690, doi: 10.1109/ICCONS.2018.8662957.

[3] A. Iyengar, G. Kalpana, S. Kalyankumar and S. GunaNandhini, "Integrated SPAM detection for multilingual emails," 2017 International Conference on Information Communication and Embedded Systems (ICICES), Chennai, India, 2017, pp. 1-4, doi: 10.1109/ICICES.2017.8070784.

[4] S. Suryawanshi, A. Goswami and P. Patil, "Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers," 2019 IEEE 9th International Conference on Advanced Computing (IACC), Tiruchirappalli, India, 2019, pp. 69-74, doi: 10.1109/IACC48062.2019.8971582.

[5] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti and M. Alazab, "A Comprehensive Survey for Intelligent Spam Email Detection," in IEEE Access, vol. 7, pp. 168261-168295, 2019, doi: 10.1109/ACCESS.2019.2954791.

[6] M. R. Islam, M. U. Chowdhury and Wanlei Zhou, "An Innovative Spam Filtering Model Based on Support Vector Machine," International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), Vienna, Austria, 2005, pp. 348-353, doi: 10.1109/CIMCA.2005.1631493.

[7] S. Gibson, B. Issac, L. Zhang and S. M. Jacob, "Detecting Spam Email with Machine Learning Optimized with Bio-Inspired Metaheuristic Algorithms," in IEEE Access, vol. 8, pp. 187914- 187932,2020,doi:10.1109/ACCESS.2020.303075

[8] A. Karim, S. Azam, B. Shanmugam and K. Kannoorpatti, "Efficient Clustering of Emails into Spam and Ham: The Foundational Study of a Comprehensive Unsupervised Framework," in IEEE Access, vol. 8, pp. 154759-154788, 2020, doi: 10.1109/ACCESS.2020.3017082.

[9] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed and M. A. Al-Garadi, "Email Classification Research Trends: Review and Open Issues," in IEEE Access, vol. 5, pp. 9044-9064, 2017, doi: 10.1109/ACCESS.2017.2702187.

[10] S. A. A. Ghaleb et al., "Feature Selection by Multiobjective Optimization: Application to Spam Detection System by Neural Networks and Grasshopper Optimization Algorithm," in IEEE Access, vol. 10, pp. 98475-98489, 2022, doi: 10.1109/ACCESS.2022.3204593.

[11] G. Al-Rawashdeh, R. Mamat and N. Hafhizah Binti Abd Rahim, "Hybrid Water Cycle Optimization Algorithm with Simulated Annealing for Spam E-mail Detection," in IEEE Access, vol. 7, pp. 143721-143734, 2019, doi:10.1109/ACCESS.2019.2944089.

[12] M. -A. Oveis-Gharan and K. Raahemifar, "Multiple classifications for detecting Spam email by novel consultation algorithm," 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), Toronto, ON, Canada, 2014, pp.1-5,doi: 10.1109/CCECE.2014.6901141.

[13] A. Subasi, S. Alzahrani, A. Aljuhani and M. Aljedani, "Comparison of Decision Tree Algorithms for Spam E-mail Filtering," 2018 1st International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, 2018, pp. 1-5, doi: 10.1109/CAIS.2018.8442016.

[14] A. Karim, S. Azam, B. Shanmugam and K. Kannoorpatti, "An Unsupervised Approach for Content-Based Clustering of Emails into Spam and Ham Through Multiangular Feature Formulation," in IEEE Access, vol. 9, pp. 135186-135209, 2021, doi: 10.1109/ACCESS.2021.3116128.