International Journal of Engineering and Management Research
e-ISSN: 2250-0758 | p-ISSN: 2394-6962
https://ijemr.vandanapublications.com

Peer Reviewed & Refereed Journal
Volume-13, Issue-6 (December 2023)
https://doi.org/10.31033/ijemr.13.6.7

Enhancing Email Communication Security through Hierarchical Machine Learning Models

Neha Mangesh Dandvekar
Computer Science & Engineering, Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal, INDIA

Corresponding Author: dandvekarneha@gmail.com

Received: 13-11-2023 Revised: 28-11-2023 Accepted: 13-12-2023

ABSTRACT

With the exponential growth of digital communication, the menace of spam emails has become a pervasive issue, threatening the efficiency and security of communication channels. To improve social media security, the detection and control of spam text are essential. This paper presents a detailed survey on the latest developments in spam text detection and classification in social media. The various techniques involved in spam detection and classification involving Machine Learning, Deep Learning, and text-based approaches are discussed in this paper. We also present the challenges encountered in the identification of spam with its control mechanisms and datasets used in existing works involving spam detection. This paper presents a comprehensive approach to designing and implementing an advanced spam detection system that leverages the power of machine learning and Natural Language processing(NLP)techniques.

Our focus in this article is to detect any deceptive text reviews. In order to achieve that we have worked with both labelled and unlabelled data and proposed Techniques such as Tokenization Stemming, Lemmatization, Tf-idf Vectorization and ML Algorithm such as Naïve Bayes, Random Forest, KNN, Support Vector Machine.

***Keywords*--** SMS, Spam, Machine Learning, NLP, Tokenization, Stemming, Lemmatization, Tf-idf Vectorization, Naïve Bayes, Random Forest, KNN, Support Vector Machine

I. INTRODUCTION

Email communication is an integral part of modern communication systems, playing a crucial role in both personal and professional spheres. However, the proliferation of email spam poses significant challenges to users, organizations, and email service providers. This research focuses on the development and implementation of an advanced email spam detection system using state-of-the-art machine learning techniques

The proposed system leverages a diverse set of features extracted from email content, headers, and sender information. Machine Learning Algorithm & Natural Language Processing (NLP) algorithms are employed to analyse the textual content of emails, while metadata features contribute to a holistic understanding of the email context. The model is trained on a large dataset comprising both spam and non-spam emails to ensure robust generalization.

The primary goal of this paper is to provide a strong and comprehensive comparative study of current research on detecting review spam using various machine learning techniques and to devise methodology for conducting further investigation.

II. LITERATURE REVIEW

The systematic literature review provides the answers to specific research questions, whereas the general survey paper gives a broad idea about the domain. The key objective of this research is to identify the best available feature extraction techniques, and present different existing models for spam review detection and available parameters to analyse these models. The process of SLR helps to determine different studies available in the domain of spam review detection and answer different research questions. The distinct phases of the systematic literature review are shown in **Figure 1**. Preceding the study, the researchers discuss how different steps are performed in each phase of SLR.

Figure 1: Necessity of the SLR

It is necessary to collect the best evidence from the existing literature. The SLR process provides the best techniques to collect and analyse evidence from primary studies. It also addresses the importance of the different methods of each research question. Following a search string is done to confirm that there exists no similar literature in this domain.

((‘Spam’ or ‘Fake’ or ‘Shill’ or ‘Opinion Spam’ or ‘Spammer’ or ‘Social Spam’) AND (‘Reviews’ or

Peer Reviewed & Refereed Journal
Volume-13, Issue-6 (December 2023)
https://doi.org/10.31033/ijemr.13.6.7

'Comments' or 'Online Comments') AND ('Detection' or 'Finding') AND ('Technique' or 'Method') AND ('Systematic Overview' or 'Systematic Review' or 'Research review'))

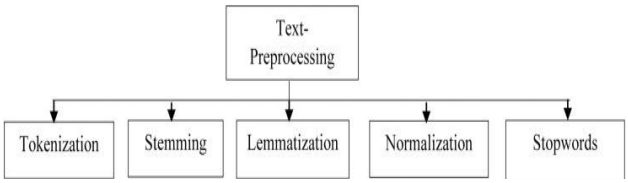
The studies are selected based on their title, abstract, and conclusion. The result of this search depicted that there is no SLR having the same scope.

1. SMS Spam Detection using Machine Learning Approach Abhishek Patel#1 , Priya Jhariya*2 , Sudalagunta Bharath 3, Ankita wadhawan#4 Computer Science Engineering Department Lovely Professional University Phagwara Punjab In this technical paper we came to know about the use of implements such as prison cell phone has extended, Short Message Service (SMS) has developed into a multi-billion dollar industry. Concurrently, a reduction in the expenditure of notifying managements has transported around expansion in impulsive business elevations (spams) being transported off prison cell phones..
2. Mobile SMS Spam Detection using Machine Learning Techniques Samadhan Nagre Dept of Computer Science & IT Dr. B.A.M. University Aurangabad In this paper we got to know about types of SMS documentation and different Detection System and There be some documented SMS spam detection system consume some trial supplementary than SMS spam detection such as incomplete communication size use of local and shortcut words and incomplete slogan information

III. METHODOLOGY

The drive of this evaluation daily is to review and disapprovingly study the numerous mechanism knowledge actions and governments employed in SMS junk mail discovery. We purpose to deliver a complete sympathetic of the present approaches, their fortes, dimness, and possible parts for progress. A research dataset be wanted intended for a number type of mechanism knowledge organization systems. Result of the mechanism knowledge procedures depend on the dataset since a effect junk discovery procedures be able to run without a dataset. within we established dissimilar openly available dataset apply in different studies. link of the dataset and a quantity of statistics such as total number of SMS number of spam and ham messages are shown in table. Result Analysis: A initial we physically searched on top of Google using the topic spam detection to increase an impression in spam detection field, It resulted inside a lot of SMS spam detection connected papers. some resources strength not contain been published in a straight line. One more danger be a number of resources are not available intended for community use

1. Pre-Processing



- ***Removing Stop Words or Punctuation***

Generally, the review text contains unnecessary words like "is", "the"," and", "a". These words are not helpful in detecting spam reviews, therefore, it is better to remove them

- ***Part of speech tagging***

This basically involves tagging word features with parts of speech Moreover, the relationship with the adjacent and related words in a review text is also tagged. A simplified form of POS tagging is the identification of words as nouns, verbs, adjectives, adverbs, etc.

- ***Stemming Word***

A stemming algorithm converts different forms of the word into a single recognized form. For instance, considering the words "works", "working", and "worked" as instances of the word work. Stemming must be applied to the review text before tokenizing it.

2. Tokenization

In this method, single words or group of words are used as features. This linguistic technique is called uni-gram when one word is selected, bi-gram when two words are selected, tri-gram when three words are selected, and so on. This technique is called n-gram

3. Transformation

Document term matrix is used to represent tokens generated by the n-gram model in the form of a sparse matrix. A sparse matrix defines the frequency of terms or tokens in the collection of the reviews. It was observed by the literature review that most of the researchers use the following two techniques for transformation.

- ***Term Frequency and Inverse Document Frequency (TF-IDF)***

TF-IDF technique is intended to consider how significant a word or token is to a document in a collection of the corpus. The TF-IDF value increases proportionally to the number of times a token occurs in the document but is often decreases by the occurrence of the word in the corpus, which helps to adjust for the fact that few words occur more often generally. Nowadays, TF-IDF is one of the most popular term-weighting schemes and provides better results than a simple count technique.

Decision Trees

A decision hierarchy be a decision support instrument that use a hierarchy similar to or classical of choices then their likely penalty, counting possibility of occasion outcomes. A choice sapling can be rummage-

Peer Reviewed & Refereed Journal
Volume-13, Issue-6 (December 2023)
https://doi.org/10.31033/ijemr.13.6.7

sale to make choice towards whether a fresh message is junk or overact.

Logistic Regression

A logistic regression remain a predictive investigation. Logistic reversion remain second hand toward explain information and to explain the association flanked by single reliant binary changeable and single or additional hypothetical ordinal, intermission before percentage-level self-governing variable quantity. after stretch toward period logistic reversion remain hard to understand, the intellects data tool without difficulty allows you to demean or the investigation, after that in simple English interprets the production.

Random Forest

Random Forest is a emblem term for an communal of choice plants. In random forest we assortment of choice plants. Toward categorize a original thing based on qualities respectively sapling elections aimed at that session. The forestry indicates the organization consuming the greatest votes finished all the foliage in the forestry.

Naive Bayes

Since 1998, Naive Bayes has been used in supervised machine learning to identify spam [20]. It primarily relies on the chance to differentiate between different entities based on predefined characteristics. Naive Bayes senses a word or event that happened in a previous context and calculates the likelihood of that word or event occurring again in the future [21]. For example, if a word appears in a spam e-mail but not in a ham e-mail, the algorithm would most likely classify it as spam.

IV. CONCLUSION

The aims and objectives of the project, which achieved throughout the course, defined at the very first stage of the process. To collect all the information, the research work involved a careful study on the different filtering algorithms and existing anti-spam tools. These largescale research papers and existing software programs are one of the sources of inspiration behind this project work. The whole project was divided into several iterations. Each iteration was completed by completing four phases: inception, where the idea of work was identified; elaboration, where architecture of the part of the system is designed; construction, where existing code is implemented; transition, where the developed part of the project is validated. However, there are still some parts that can be improved: for example, adding additional filtering techniques or changing aspects of the existing ones. The changes such as incrementing or decrementing the number of interesting words of the message and reorganizing the formula for calculating interesting rate can be done later.

REFERENCES

[1] SP. Rajamohana, Dr. K. Umamaheswari, M. Dharani & R. Vedackshya. (2017). A survey on online review spam detection techniques. In: *International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT)*.

[2] Minqing Hu & Bing Liu. (2004). Mining and summarizing customer reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[3] Camponovo G & Cerutti D. The spam issue in mobile business: A comparative regulatory overview. *Proc. of 3rd Int. Conf. Mobile Bus.*, pp. 1-17.

[4] Cleff E.B. Privacy issues in mobile advertising. *Int. Rev. Law Computer. Technol.*, *21*, 225-236.

[5] Fu J, Lin P & Lee S. Detecting spamming activities in a campus network using incremental learning. *J. Network .Computer. Appl.*, *43*, 56-65.

[6] N. Kumar, S. Sonowal & Nishant. (2020). Email spam detection using machine learning algorithms. In: *Proceedings of the 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 108–113.

[7] G. Jain, M. Sharma & B. Agarwal. (2019). Optimizing semantic lstm for spam detection. *International Journal of Information Technology*, *11*(2), 239–250.