DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SIDDAGANGA INSTITUTE OF TECHNOLOGY, TUMKUR-572103

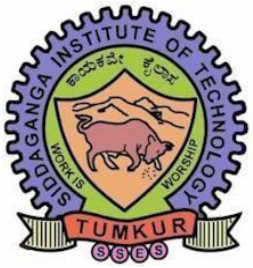
MINI PROJECT SYNOPSIS 2020-21

Spam Detection

1. **RISHIKA KUMARI(1SI18CS087)**
2. **RIYA JAISWAL(1SI18CS093)**
3. **RIYUKSH SHODA(1SI18CS094)**

Under the guidance of

H K Vedamurthy
Assistant Professor



Department of Computer Science and Engineering

Siddaganga Institute of Technology, Tumakuru – 572103

(An Autonomous Institution, Affiliated to VTU, Belagavi& Recognized by AICTE, New Delhi)

2020 -2021

INTRODUCTION

All mails have a common structure i.e. subject of the email and the body of the mail. A typical spam mail can be classified by filtering its content. The proceess of spam mail detection is based on the assumption that the content of the spam mail is different than the legitmate mail.The proposed spam mail detection system is inspired from the effectiveness of machine learning approach.

In mail spam detection, initially data is collected. The data is raw data and unstructured in nature in nature.In order to reduce the computation and to obtain accurate results, email data needds to be processed. The data is pre-proccessed by removing stop words, stemming and word tokenization is also performed to acquire valuable information. This step reduces thed imensionality of data and featured in form of bag of words are then extracted.For the classification a bagged hybrid approach is used in order to make the classification stronger and more accurate. The dataset is randomly divided into classification algorithm.

OBJECTIVES

- The main objective of this project is to aware the user regarding fake email and relevant emails.
- To also classify the mail whether mail is spam or not.

MOTIVATION

With the influx of technological advancements and the increased simplicity in communication, especially through emails, the upsurge in the volume of unsolicited bulk emails (UBEs) has become a severe threat to global security and economy. Spam emails not only waste users time, but also consume a lot of network bandwidth, and may also include malware as executable files. Thus there is an intrinsic need for the development of more robust and dependable UBE filters that facilitate automatic detection of such emails.

LITERATURE SURVEY

FIRST REPORT:

Survey on web spam detection

By Jiawei Han and Nikita Spirin

Search engines became a well know place to start information acquisition on the Web. Though due to web spam phenomenon, search results are not always as good as desired. Moreover, spam evolves that makes the problem of providing high quality search even more challenging. Over the last decade research on adversarial information retrieval has gained a lot of interest both from academia and industry. In this paper we present a systematic review of web spam detection techniques with the focus on algorithms and underlying principles. We categorize all existing algorithms into three categories based on the type of information they use: content-based methods, link-based methods, and methods based on non-traditional data such as user behaviour, clicks, HTTP sessions. In turn, we perform a subcategorization of link-based category into five groups based on ideas and principles used: labels propagation, link pruning and reweighting, labels refinement, graph regularization, and feature-based. We also define the concept of web spam numerically and provide a brief survey on various spam forms. Finally, we summarize the observations and underlying principles applied for web spam detection.

SECOND REPORT:

Spam Review Detection Techniques: A Systematic Literature Review

by N Hussain, Hamid Turab Mirza, Ghulam Rasool, Ibar Hussain and Mohammad Kaleem.

Online reviews about the purchase of products or services provided have become the main source of users opinions. In order to gain profit or fame, usually spam reviews are written to promote or demote a few target products or services. This practice is known as review spamming. In this study, a comprehensive review of existing studies on spam review detection using the Systematic Literature Review (SLR) approach. This literature review identified two major feature extraction techniques and two different approaches to review spam detection. In addition, this study has identified different performance metrics that are commonly used to evaluate the accuracy of the review spam detection models. Lastly, this work presents an overall discussion about different feature extraction approaches from review datasets, the proposed taxonomy of spam review detection approaches, evaluation measures, and publicly available review datasets. The feature's extraction depends upon the review dataset, and the accuracy of review spam detection methods is dependent upon the selection of the feature engineering approach. Therefore, for the successful implementation of the spam review detection model and to achieve better accuracy, these factors are required to be considered in accordance with each other. To the best of the researchers' knowledge, this is the first comprehensive review of existing studies in the domain of spam review detection using SLR process.

Review spam detection via temporal pattern discovery by SihongXie ,Guan WangShuyang and Lin P S Yu

Online reviews play a crucial role in today's electronic commerce. It is desirable for a customer to read reviews of products or stores before making the decision of what or from where to buy. Due to the pervasive spam reviews, customers can be misled to buy low-quality products, while decent stores can be defamed by malicious reviews. We call this problem singleton review spam detection.

To address this problem, we observe that the normal reviewers' arrival pattern is stable and uncorrelated to their rating pattern temporally. In contrast, spam attacks are usually bursty and either positively or negatively correlated to the rating. Thus, we propose to detect such attacks via unusually correlated temporal patterns. We identify and construct multidimensional time series based on aggregate statistics, in order to depict and mine such correlations. In this way, the singleton review spam detection problem is mapped to a abnormally correlated pattern detection problem. We propose a hierarchical algorithm to robustly detect the time windows where such attacks are likely to have happened. The algorithm also pinpoints such windows in different time resolutions to facilitate faster human inspection. Experimental results show that the proposed method is effective in detecting singleton review attacks. We discover that singleton review is a significant source of spam reviews and largely affects the ratings of online stores.

THIRD REPORT:

Ensemble based spam detection in social IoT using probabilistic and data structures by Amritpal Singh and Shalini Batra.

A social approach can be used for the Internet of Things (IoT) to connect large number of objects in social networks like Twitter, Facebook, Instagram and many more. Social networks within the IoT domain have simplified the task of dynamic discovery of services and information. Detecting spam in social media, especially when massive data flows continuously and large number of attributes are associated with it, is a daunting task which requires lot of technical insight. This paper proposes a semi-supervised technique for spam detection in Twitter by employing ensemble based framework comprising of four classifiers. The framework is based on usage of Probabilistic Data Structures (PDS) like Quotient Filter (QF) to query the URL database, spam users, spam words databases and Locality Sensitive Hashing (LSH) for similarity search, as classifiers in various stages which provide fast results with less computational effort.
Performance of the framework has been evaluated by comparative analysis of PDS with the similar data structures and through the standard evaluation parameters which include precision, recall and F-score.

FOURTH REPORT:

Effectively Detecting Content Spam on the Web Using Topical Diversity Measures. By Cailing Dong and Bin Zhou.

Recent studies about web spam detection have utilized various content-based and link-based features to construct a spam classification model. In this paper, we conduct a thorough analysis of content spam on the web using topic models and propose several novel topical diversity measures for content spam detection. We adopt the web spam benchmark data set WEBSPAM-UK2007 for evaluation, and the experimental results verify that by integrating our topical diversity measures the performance of the state-of-the-art web spam detection methods can be greatly improved. In addition, comparing to existing features for training spam classification models, our topical diversity measures can achieve high spam detection performance using small set of training data. In personalized web spam detection, the training data are typically small.

Our finding makes personalized web spam detection highly achievable. We develop an efficient and effective regression model using topical diversity measures for personalized web spam detection, and present some promising results obtained from an empirical study.

FIFTH REPORT:

Fuzzy Improved Decision Tree Approach for Outlier Detection in SMS. By Priyanka Maan and Meghna Sharma.

Spam is one of the serious problems faced by internet community globally. Spam Detection is a critical issue in business world. In this paper an intelligent three stage model is presented to perform the spam inclusive outlier identification. The SMS textual dataset is taken as input and than its filtration is done. After that this textual information is converted to the statistical information using fuzzy and assign the weights to dataset. The decision tree algorithm is than applied on this fuzzy weighed dataset to classify the dataset. This algorithm is defined to separate the spam and non spam data values. A comparison of existing Bayesian and proposed Fuzzy based decision tree approach is done. The results shows that the recognition rate is improved using the proposed approach. The work is implemented in weka integrated java environment.

SIXTH REPORT:

Feature Subset Selection Using Binary Quantum Particle Swarm Optimization for Spam Detection System. By Behjat , Amir Rajabi ,Mustapha ,Aida ,Nezamabadi-Pour,

Hossein ,Sulaiman , Md. Nasir , Mustapha, Norwati

E-mail is efficient and common communication method these days, but flooding spam or unsolicited e-mail messages have become uncontrollable. Most spams benefit from the commercial advertising. From the literature review only the accuracy identify the classification process in spam detection. Infrequently, false positive is identified as measurement methods of detection system. Based on the proposed detection system of this study, for the first time Binary Quantum Particle Swarm Optimization (BQPSO) as feature selection method decrease the number of irrelevant features in order to increase classifier performance and decrease dimensionality that influence reliability of detection system. The Multi-Layer Perceptron (MLP) classifier is applied in this research to detect spam emails based on selected relevant features. The experiments are showed on two datasets, namely LingSpam and Spam-Assassin to indicate BQPSO based on MLP classifier not only reduce high dimensionality but achieves the accuracy near to 100% with less false positive rate in spam detection system.

METHODOLOGY

Method:-Naive Bayes Spam Filtering

Naive Bayes Classifier are a popular statistical technique of e-mail filtering. They typically use bag of words features to identify spam e-mail, an approach commonly used in text classification.

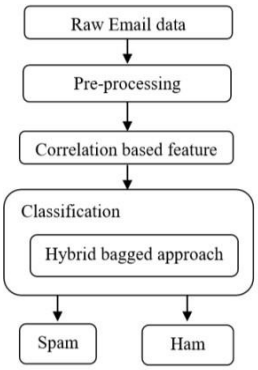


Figure. 1 Basic process for email filtering

TOOLS AND TECHNOLOGIES

- Python Language-**Python** is an interpreted, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

- Jupyter Notebook by Anaconda- **Anaconda** is a conditional free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment.

EXPECTED OUTCOMES

- We would be able to distinguish between the mails that are spam and the rest.
- We would be able to help the user to distinguish between fake emails and relevant emails.

CONCLUSION

In this study, we reviewed machine learning approach in the field of spam filtering. A Naive Bayes algorithm been applied for classification of messages as either spam or ham is provided. The attempts are made to solving the problem of spam through the use of machine learning classifiers is discussed. The evolution of spam messages over the years to evade filters is examined. The basic architecture of email spam filter and the processes involved in filtering spam emails are looked into. The challenges of the machine learning algorithms in efficiently handling the menace of spam is pointed out using the machine learning technics . Having discussed the open problems in spam filtering, further research to enhance the effectiveness of spam filters need to be done.