OPTIMIZING SPAM FILTERING WITH MACHINE LEARNING

LITERATURE SURVEY

Literature for Spam Filtering

project would involve researching and analysing existing studies, papers, and articles on the topic to gain a thorough understanding of the current state of SMS spam classification and to identify potential areas for improvement and future research. The survey would include looking at different methods and techniques used for identifying and flagging spam messages, such as machine learning algorithms, natural language processing, and rule-based systems. It would also involve evaluating the performance and effectiveness of

these methods, as well as their limitations and challenges. Additionally, the literature survey would review the current state of SMS spam and trends in the industry, as well as any existing laws and regulations related to spam messaging. The survey would also investigate the datasets and feature representations used in previous studies, which would help to determine the best approach for the current project. Furthermore, It would be important to check the pre-processing techniques used in the research to understand how to properly clean and prepare the data for the classifier

Many SMS Spam messages detection techniques are available these days to block spam messages and filtering spam messages. Few of which are mentioned below: - Gómez Hidalgoet. al. assessed a few Bayesian based classifiers to identify mobile phone spam. In this problem, the researchers proposed the first two surely understood SMS spam datasets: the Spanish (199spam and 1,157 ham) and English (82 spam and 1,119 ham) test databases. They have tried on them various messages portrayal techniques and machine learning calculations, as far as viability. The outcomes show that Bayesian separating methods can be successfully utilized to group SMS spam

Of all the different medium communication, email is extremely important medium now a days. It has been used widely for formal online communication. It can be accessed from any part of the world just with the help of internet connectivity. According to D Tschabitscher, number ofactive email accounts was 5 billion in 2017 and is increasing exponentially. He also stated that, everyday more than 270 billion Emails are exchanged, but the worst part of that is, out of that approximately 57 % emails are of no use as they are spam emails. Spam emails are creating a serious problem to the user as spammers flood the user’s system with spam emails which results in storage problem, consumption of bandwidth and leads to decrease in performance of system. Spam emails are called as junk emails or unsolicited message which is set by spammer through email. To make the email more secure and effective, appropriate email filtering is essential. Several types of researches have been done on email filtering, some acquired good accuracy but theprogress is needed in this field. In order to avoid detection,spammers came with a new approach for sending spams to other users. It is included in the advertisements as the part ofan embedded image file attachment in the form of.gif, .jpg, .png, etc. rather than body of the emails, hence by passing text-based spam filtering techniques. As we know that there are many techniques already there for email spam detection, our project aims for questing and analyzing the efficiency of the vital technique used for spam email detection from images and PDFs using Multinomial Naive Bayes' algorithm.

The author has worked with different machine learning algorithms for email classification such as Neural Network (NN), Support Vector Machine (SVM), J48 Decision Tree based classifier, Naïve Bayes. The dataset used by the author was Spam Base dataset. In this paper work, the author didn't mention advantages and disadvantages of anyalgorithm.

[G. Mujtaba] [L. Shuib] [R. G. Raj] [N. Majeed] [M. A.Al- Garadi] (2017) [2]Proposed the basic three steps which are common in every classification process. The first step is pre-processing in which the given text is converted into tokens and this step is also used for removal of stop words. The second step is learning process and, in this feature, set is built which is very much necessary for the classification of emails

REFERENCES

[1] S. K. Tuteja, "Classification Algorithms for Email Spam Filtering", 2016.

[2] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, "Email Classification ResearchTrends: Review and Open Issues", 2017.

[3] S. Ajaz, M. T. Nafis, and V. Sharma, "Spam Mail Detection Using Hybrid Secure Hash Based Naïve Classifier, 2017.