VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

Department of Computer Engineering

SMS SPAM DETECTION

In partial fulfillment of the Final Year, Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai Academic Year 2020-2021.

Submitted by

Anish Adnani D17B 01

Nihal Bhandary D17B 10

Supriya Patil D17B 50

Project Mentor

Prof. Vidya Zope
Prof. Ashvini Gaikwad

INTRODUCTION

SMS is a text communication platform that allows mobile phone users to exchange short text messages (usually less than 160 seven-bit characters). It is the most widely used data application with an estimated 3.5 billion active users. Although we have seen a decrease in text spams in early 2010s due to people switching over to other messaging applications, SMS spams still remain a persistent problem as OTP messages are now being used for a security link to F2A standards. SMS have become more relevant than ever. Phishing scams, SMS bombs, Virus embedded SMS have shown an increase in the past couple of years. With some of them being capable of evading existing spam filters. As previously mentioned with the decrease in popularity of SMS, the focus for maintaining and updating existing spam filters have been changed to prioritize spam filters for email, other messaging applications,etc.

Research Papers Referred

- Shafi'i Muhammad Abdulhamid, Muhammad Shafie Abd Latiff, Haruna Chiroma, Oluwafemi Osho, GaddafiAbdul-Salaam, Abdul - Salaam, Adamu I. Abubakar and Tutut Herawan. "A Review on Mobile SMS Spam Filtering Techniques." IEEE Access 2017. Doi: 10.1109/ACCESS.2017.2666785. INSPEC Accession Number: 17124287.The system used in this paper has analyzed and studied the relative strengths of various machine learning algorithms in order to detect spam messages. Open dataset is used and prepared two datasets for testing and validation purposes. Accuracy in detecting spam messages was the first priority in ranking these algorithms. In order to evaluate or determine the accuracy of the mobile SMS spam filtering techniques, certain performance evaluation metrics like precision, recall, accuracy, F1 score were applied to the selected paper. The number of support vectors (SV) is directly proportional to the size of the training dataset which forces SVMs to use unnecessary basis functions. The Bayesian network is still being used, despite the availability of improved and more powerful algorithms.

- Ashraf M. Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. Multinomial Naive Bayes for Text Categorization Revisited. Australian joint conference on Artificial Intelligence pp 488-499. ISBN 978-3-540-30549-1 Doi: https://doi.org/10.1007/978-3-540-30549-1_43. The system used in this paper, presents empirical results for several versions of the multinomial naive Bayes classifier on four text categorization prob-lems, and a way of improving it using locally weighted learning. The paper shows how the performance of multinomial naive Bayes can be improved using locally

weighted learning. However,the overall conclusion of our paper is that support vector machines are still the method of choice if the aim is to maximize accuracy.

- “Text normalization and semantic indexing to enhance Instant Messaging and SMS spam filtering.” Tiago A. Almeida, Tiago P. Silva, Igor Santos and Jose M. Gomez Hidalgo. Doi: https://doi.org/10.1016/j.knosys.2016.05.001. The system present in this paper focuses on preprocessing: text normalization and semantic indexing which combination can prove to be best suited for SMS spam filtering. It proposes and then evaluates a method to normalize and expand original short and messy text messages in order to acquire better attributes and enhance the classification performance. Two main issues make the application of established classi-fication algorithms difficult for this specific field of research: the low number of features that can be extracted per message and the fact that messages are filled with slangs and idioms.

- Inwhee Joe and Hyetaek Shim. “An SMS Spam Filtering System Using Support Vector Machine.” International Conference on Future Generation Information Technology. pp 577-584. ISBN 978-3-642-17569-5 Doi:https://doi.org/10.1007/978-3-642-17569-5_56. This paper focuses on SMS filtering applications using support vector machines. They extract feature vectors, then train the learning component, which is then used to filter the spam. In this they are using three components The first is a feature vector generator component that generates feature vectors after training. The second one is the SVM learner component using the generated feature vector. The last one is the spam filtering component to categorize spam

messages using the completed classifier. Drawback of the system is, it is Developed in Windows environment its accuracy reaches up to 75% which is acceptable but not as high as that in the case of Multinomial Naive Bayes.

- “A Discrete Hidden Markov Model for SMS Spam Detection.” Tian Xia and Xuemin Chen. doi: https://doi.org/10.3390/app10145011. The system in this paper proposed a new method based on the discrete hidden Markov model (HMM) to use the word order information and to solve the low term frequency issue in SMS spam detection. The popularly adopted SMS spam dataset from the UCI machine learning repository is used for performance analysis of the proposed HMM method. Dataset used was not properly distributed with more number of ham (no spam) classes in the dataset.

References	Abstract	Keywords	Our Findings
A Review on Mobile SMS Spam Filtering Techniques	Relative strengths of various machine learning algorithms in order to detect spam messages. The open dataset is used and prepared two datasets for testing and validation purposes. Accuracy in detecting spam messages was the first priority in ranking these algorithms	Spam detection, SMS, Accuracy	In order to evaluate or determine the accuracy of the mobile SMS spam filtering techniques, certain performance evaluation metrics like precision, recall, accuracy, F1 score were applied to the selected paper.
Multinomial Naive Bayes for Text Categorization Revisited	Empirical results for several versions of the multinomial naive Bayes classifier on four text categorization problems, and a way of improving it using locally weighted learning.	Review, spam, mobile SMS, access layer, service provider layer voices.	The paper shows how the performance of multinomial naive Bayes can be improved using locally weighted learning. However, the overall conclusion of our paper is that support vector machines are still the method of choice if the aim is to maximize accuracy.
Text normalization and semantic indexing to enhance Instant Messaging and SMS spam	Preprocessing: text normalization and semantic indexing which combination can prove to be best suited for SMS spam filtering	Normalization, SMS, Text categorization	Two main issues make the application of established classification algorithms difficult for this specific field of research: the low number of features that can be extracted per message and the fact

filtering			that messages are filled with slang and idioms.
An SMS Spam Filtering System Using Support Vector Machine	Extract feature vectors, then train the learning component, which is then used to filter the spam	Spam filtering, system short messaging, service support vector machine, thesaurus.	It is Developed in a Windows environment its accuracy reaches up to 75% which is acceptable but not as high as that in the case of Multinomial Naive Bayes