Abstract

Nowadays communication plays a major role in everything be it professional or personal. Email communication service is being used extensively because of its free use services, low-cost operations, accessibility, and popularity. Emails have one major security flaw that is anyone can send an email to anyone just by getting their unique user id. This security flaw is being exploited by some businesses and ill-motivated persons for advertising, phishing, malicious purposes, and finally fraud. This produces a kind of email category called SPAM.

Spam refers to any email that contains an advertisement, unrelated and frequent emails. These emails are increasing day by day in numbers. Studies show that around 55 percent of all emails are some kind of spam. A lot of effort is being put into this by service providers. Spam is evolving by changing the obvious markers of detection. Moreover, the spam detection of service providers can never be aggressive with classification because it may cause potential information loss to incase of a misclassification.

To tackle this problem we present a new and efficient method to detect spam using machine learning and natural language processing. A tool that can detect and classify spam. In addition to that, it also provides information regarding the text provided in a quick view format for user convenience.

TABLE OF CONTENTS

List of Figures

List of Tables

1. Introduction

Today, Spam has become a major problem in communication over internet. It has been accounted that around 55% of all emails are reported as spam and the number has been growing steadily. Spam which is also known as unsolicited bulk email has led to the increasing use of email as email provides the perfect ways to send the unwanted advertisement or junk newsgroup posting at no cost for the sender. This chances has been extensively exploited by irresponsible organizations and resulting to clutter the mail boxes of millions of people all around the world.

Spam has been a major concern given the offensive content of messages, spam is a waste of time. End user is at risk of deleting legitimate mail by mistake. Moreover, spam also impacted the economical which led some countries to adopt legislation.

Text classification is used to determine the path of incoming mail/message either into inbox or straight to spam folder. It is the process of assigning categories to text according to its content. It is used to organized, structures and categorize text. It can be done either manually or automatically. Machine learning automatically classifies the text in a much faster way than manual technique. Machine learning uses pre-labelled text to learn the different associations between pieces of text and it output. It used feature extraction to transform each text to numerical representation in form of vector which represents the frequency of word in predefined dictionary.

Text classification is important to structure the unstructured and messy nature of text such as documents and spam messages in a cost-effective way. Machine learning can make more accurate precisions in real-time and help to improve the manual slow process to much better and faster analysing big data. It is important especially to a company to analyse text data, help inform business decisions and even automate business processes.

In this project, machine learning techniques are used to detect the spam message of a mail. Machine learning is where computers can learn to do something without the need to explicitly program them for the task.

It uses data and produce a program to perform a task such as classification. Compared to knowledge engineering, machine learning techniques require messages that have been successfully pre-classified. The pre-classified messages make the training dataset which will be used to fit the learning algorithm to the model in machine learning studio.

A combination of algorithms are used to learn the classification rules from messages. These algorithms are used for classification of objects of different classes. These algorithms are provided with pre labelled data and an unknown text. After learning from the prelabelled data each of these algorithms predict which class the unknown text may belong to and the category predicted by majority is considered as final.

2. Literature Review

2.1 Introduction

This chapter discusses about the literature review for machine learning classifier that being used in previous researches and projects. It is not about information gathering but it summarize the prior research that related to this project. It involves the process of searching, reading, analysing, summarising and evaluating the reading materials based on the project.

A lot of research has been done on spam detection using machine learning. But due to the evolvement of spam and development of various technologies the proposed methods are not dependable. Natural language processing is one of the lesser known fields in machine learning and it reflects here with comparatively less work present.

2.2 Related work

Spam classification is a problem that is neither new nor simple. A lot of research has been done and several effective methods have been proposed.

i. M. RAZA, N. D. Jayasinghe, and M. M. A. Muslam have analyzed various techniques for spam classification and concluded that naïve Bayes and support vector machines have higher accuracy than the rest, around 91% consistently [1].

ii. S. Gadde, A. Lakshmanarao, and S. Satyanarayana in their paper on spam detection concluded that the LSTM system resulted in higher accuracy of 98%[2].

iii. P. Sethi, V. Bhandari, and B. Kohli concluded that machine learning algorithms perform differently depending on the presence of different attributes [3].

iv. H. Karamollaoglu, İ. A. Dogru, and M. Dorterler performed spam classification on Turkish messages and emails using both naïve Bayes classification algorithms and support vector machines and concluded that the accuracies of both models measured around 90% [4].

v. P. Navaney, G. Dubey, and A. Rana compared the efficiency of the SVM, naïve Bayes, and entropy method and the SVM had the highest accuracy (97.5%) compared to the other two models [5].

vi. S. Nandhini and J. Marseline K.S in their paper on the best model for spam detection it is concluded that random forest algorithm beats others in accuracy and KNN in building time [6].

vii. S. O. Olatunji concluded in her paper that while SVM outperforms ELM in terms of accuracy, the ELM beats the SVM in terms of speed [7].

viii. M. Gupta, A. Bakliwal, S. Agarwal, and P. Mehndiratta studied classical machine learning classifiers and concluded that convolutional neural network outperforms the classical machine learning methods by a small margin but take more time for classification [8].

ix. N. Kumar, S. Sonowal, and Nishant, in their paper, published that naïve Bayes algorithm is best but has class conditional limitations [9].

x. T. Toma, S. Hassan, and M. Arifuzzaman studied various types of naïve Bayes algorithms and proved that the multinomial naïve Bayes classification algorithm has better accuracy than the rest with an accuracy of 98% [10].

F. Hossain, M. N. Uddin, and R. K. Halder in their study concluded that machine learning models outperform deep learning models when it comes to spam classification and ensemble models outperform individual models in terms of accuracy and precision [11].

2.3 Summary

From various studies, we can take that for various types of data various models performs better. Naïve Bayes, random forest, SVM, logistic regression are some of the most used algorithms in spam detection and classification.

3. Objectives and Scope

3.1 Problem Statement

Spammers are in continuous war with Email service providers. Email service providers implement various spam filtering methods to retain their users, and spammers are continuously changing patterns, using various embedding tricks to get through filtering. These filters can never be too aggressive because a slight misclassification may lead to important information loss for consumer. A rigid filtering method with additional reinforcements is needed to tackle this problem.

3.2 Objectives

The objectives of this project are

i. To create a ensemble algorithm for classification of spam with highest possible accuracy.

ii. To study on how to use machine learning for spam detection.

iii. To study how natural language processing techniques can be implemented in spam detection.

iv. To provide user with insights of the given text leveraging the created algorithm and NLP.

3.3 Project Scope

This project needs a coordinated scope of work.

i. Combine existing machine learning algorithms to form a better ensemble algorithm.

ii. Clean, processing and make use of the dataset for training and testing the model created.

iii. Analyse the texts and extract entities for presentation.

3.4 Limitations

This Project has certain limitations.

i. This can only predict and classify spam but not block it.

ii. Analysis can be tricky for some alphanumeric messages and it may struggle with entity detection.

iii. Since the data is reasonably large it may take a few seconds to classify and anlayse the message.

4. Experimentation and Methods

4.1 Introduction

This chapter will explain the specific details on the methodology being used to develop this project. Methodology is an important role as a guide for this project to make sure it is in the right path and working as well as plan. There is different type of methodology used in order to do spam detection and filtering. So, it is important to choose the right and suitable methodology thus it is necessary to understand the application functionality itself.

4.2 System Architecture

The application overview has been presented below and it gives a basic structure of the application.

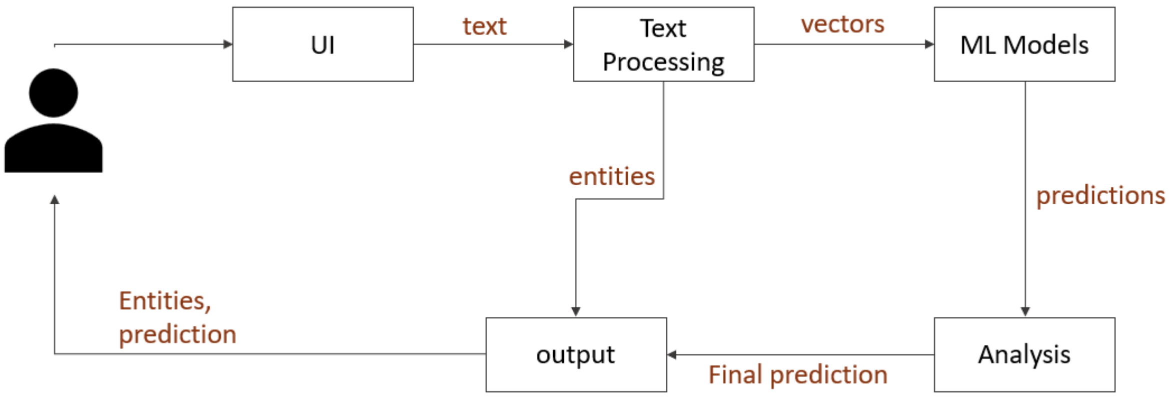

fig no. 4.1 Architecture

The UI, Text processing and ML Models are the three important modules of this project. Each Module's explanation has been given in the later sections of this chapter.

A more complicated and detailed view of architecture is presented in the workflow section.

4.3 Modules and Explanation

The Application consists of three modules.

i. UI

ii. Machine Learning

iii. Data Processing