Spam Detection Model

COMP 333: Data Analytics

Instructor: Dr. Yaser Esmaeili Salehani

Submitted by:

Matteo Robidoux
Student ID: 40282589

Raagav Prasanna
Student ID: 40282749

Date:

April 7th, 2025

Concordia University

Abstract

This study looks into the problem of finding spam in different types of input, such as SMS, email, and YouTube comments. Traditional spam detection commonly employs basic models like text vectorization. However, our study indicates that the identification of URLs embedded within messages plays a critical role in determining spam likelihood. Therefore, in addition to conventional text-based training approaches, we introduced a separate URL-based model to enhance prediction accuracy.

Furthermore, we observed key differences across input datasets. SMS messages consist solely of a body containing text, whereas emails include a subject line. Moreover, YouTube comments also often contained spam related to self-promotion, with messages similar to “Please subscribe to my channel”. Additionally, emails and comments provided unique attributes, such as email subjects and the authors of comments, which are also key pieces of data required to identify spam.

To address these variations, we developed three specialized models for SMS, email, and comment spam detection, each catering to the different datasets that we acquired. The individual model’s predictions were then combined with the output of the URL spam model whenever a URL was present. This hybrid approach significantly improved classification accuracy, achieving 97.99% accuracy for the URL model, 96.00% for SMS, 96.00% for email, and 95.33% for YouTube comments, significantly outperforming traditional methods.

1 Introduction

The rise of digital communication has led to an increase in the amount of unwanted and potentially harmful spam messages across various platforms. This includes SMS messages, emails, and content within online comment sections. Spam messages can range anywhere from harmless advertisements, to phishing attempts and even malware distribution. This emphasizes the need for effective spam detection, which is needed for ensuring the security of a user, as well as the integrity of a platform. Traditional spam detection techniques often rely solely on simple text-based models, such as a TF-IDF Vectorizer. While these methods are effective when used to identify many different spam messages given the right input dataset, they often fail to account for additional information catering to the input, leading to both false negatives as well as false positives. In short, on many occasions, messages that are not spam are marked as spam, and messages that are, are not.

One key aspect of spam identification that is not present within traditional methods is the identification of URLs. Many spam messages contain malicious links, requiring models to take these links into account. Furthermore, a message can also contain a valid url sent from a valid user, which means that messages containing urls should not be marked as spam simply because they contain a url. Instead, the url itself should be analyzed to determine whether or not it is spam. Our study demonstrates that integrating URL-based detection with text-based models significantly improves the ability to detect spam, particularly when it comes to wrongful spam detection.

Moreover, spam characteristics differ across various platforms. SMS messages primarily consist of just text, whereas emails include components like subject lines. YouTube comments, on the other hand, often involve engagement-driven spam, such as self-promotion; as well as peculiar author names that could indicate spam. These differences require the need to develop specific models for each input type, as opposed to just one model.

To address these challenges, we propose a hybrid spam detection algorithm consisting of three specialized models for SMS, email, and YouTube comments. Each model is designed to identify the unique characteristics of its respective dataset, while utilizing a separate independent model for the embedded URLs in the event they exist. When a message contains a URL, the URL model’s prediction is combined with the text model’s prediction, to determine whether the message is spam. Our results demonstrate that this approach significantly enhances detection accuracy, outperforming traditional methods.

The rest of this paper is organized as follows: Section 2 reviews related work in spam detection, and why these models are inefficient. Section 3 describes our methodology, including dataset preprocessing, model architectures, and training procedures. Section 4 presents our experimental results and evaluation metrics. Finally, Section 5 discusses key findings, limitations, and potential future improvements.

2 Related Work

Spam detection has been extensively studied across various communication platforms, including SMS, emails, and online comments. Traditional spam detection approaches, as mentioned previously, use vectorization techniques such as TF-IDF, where all of the English stop words are removed from the text and only the necessary keywords remain for training. While these methods have been successful, they often overlook some important features that are very helpful in spam classification.

A notable study by Gupta et al. follows this conventional spam detection algorithm using TF-IDF vectorization [1]. Their research applies machine learning models to the UCI SMS Spam Collection dataset [2], using text-based features extracted from SMS messages. Their results show strong performance in spam classification, using traditional approaches such as Naïve Bayes and SVM, emphasizing that this is an effective technique.

However, while training models on the same dataset [2], we identified a significant limitation in the way spam messages were labeled, especially when a URL was present in the message body. The dataset shows a strong bias, as nearly all SMS messages containing URLs are labeled as spam. This introduces a risk that models trained on this data may incorrectly generalize, leading to a high false positive rate for messages containing legitimate URLs, significantly impacting simple sms messages such as a getting sent a video link from a friend. Consequently, SMS messages may be disproportionately influenced by the presence of a URL rather than the actual content of the message.

This observation highlights a critical shortcoming of traditional text-based spam detection models, being the lack of a dedicated mechanism for evaluating URLs. Our solution resolves this issue by introducing a separate URL-based classification model, which is then used in conjunction with either the SMS message model, Email model, or YouTube comments model, to predict whether they are spam. By independently determining the likelihood of a URL being spam, this hybrid approach reduces dataset bias and ensures more balanced spam classification, particularly for messages containing URLs.

References

[1] S. D. Gupta, S. Saha, and S. K. Das, ”SMS Spam Detection Using Machine Learning,” *J. Phys. Conf. Ser.*, vol. 1797, no. 1, p. 012017, Feb. 2021, doi: 10.1088/1742-6596/1797/1/012017.

[2] UCI Machine Learning Repository, ”SMS Spam Collection Dataset,” [Online]. Available: https://www.kaggle.com/data
spam-collection-dataset. [Accessed: 2025-03-24].