Optimizing Spam Filtering Using Machine Learning

Milestone-1

Milestone 1: Define Problem / Problem Understanding

Activity 1: Specify the business problem

In today's world, email is used in almost every industry, from business to education. Emails can be categorized into two categories: ham and spam. Junk emails, also known as spam messages, are emails that have been designed to harm recipients by wasting their time, computing resources, and stealing their valuable information. It is estimated that spam emails are increasing at a rapid rate. One of the most important and prominent spam prevention techniques is filtering email. Naive Bayes, Decision Trees, Neural Networks, and Random Forests are among the methods used for this purpose by researchers. In this project, I examine the Logistic Regression machine learning model for spam filtering in email by categorizing messages into appropriate groups. This study also compares the techniques based on accuracy, precision, recall, etc. The accuracy level for this project was around 97%. Towards the end, these insights and future research directions, and challenges are outlined.

Activity 2: Business requirements

A business requirement for an SMS spam classification system would include the ability to accurately identify and flag spam messages, protect customers from unwanted or harmful messages, and comply with industry regulations and laws regarding spam messaging. Additionally, the system should be able to handle a high volume of messages, integrate with existing systems and databases, and provide reporting and analysis capabilities to track performance and improve the system over time. The system should also have an easy-to-use interface and be easy to maintain and update.

Activity 3: Literature Survey

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.

Activity 4: Social or Business Impact.

Social Impact:- it can help protect individuals from unwanted and potentially harmful messages. Spam messages can include phishing attempts, scams, and fraud, which can have serious financial and personal consequences for recipients. By accurately identifying and flagging spam messages, the system can help prevent these types of attacks and protect individuals from falling victim to them.

Business Model/Impact:- it can help protect their customers and improve their reputation. Spam messages can harm a business's reputation and lead to customer complaints and lost business. By accurately identifying and flagging spam messages, the system can help protect businesses and improve their customer's trust.