CSCE 5290: Natural Language Processing

Project Proposal

Group-2

Naveed Ahmed Mohammed - 11548614

Iliyas Ahmed Mohammed - 11550253

Mohammed Abdul Qayyoom Shaik - 11532186

Eswara Reddy Thimmapuram - 11506566

Project Title

Spam Message Classification

Goals and Objectives:

- **Motivation:**

Nearly everyone now-a-days has a smart phone that at the very least has the basic feature such as messaging. Spam messages are unsolicited text messages that are forwarded to wide group of users that are typically sent for the aim of promoting products and services without their prior permission. The goal of this study is to identify whether a given message is spam or ham from the given message by using NLP techniques and machine learning algorithms.

- **Significance:**

These days, the number of spam messages among scams has surged dramatically. These spam messages generally lure users to provide confidential and payment information by offering fraudulent and appealing deals. Using this spam detection model, we can easily identify spam messages i.e., fake messages that users are unable to recognize.

- **Objectives:**

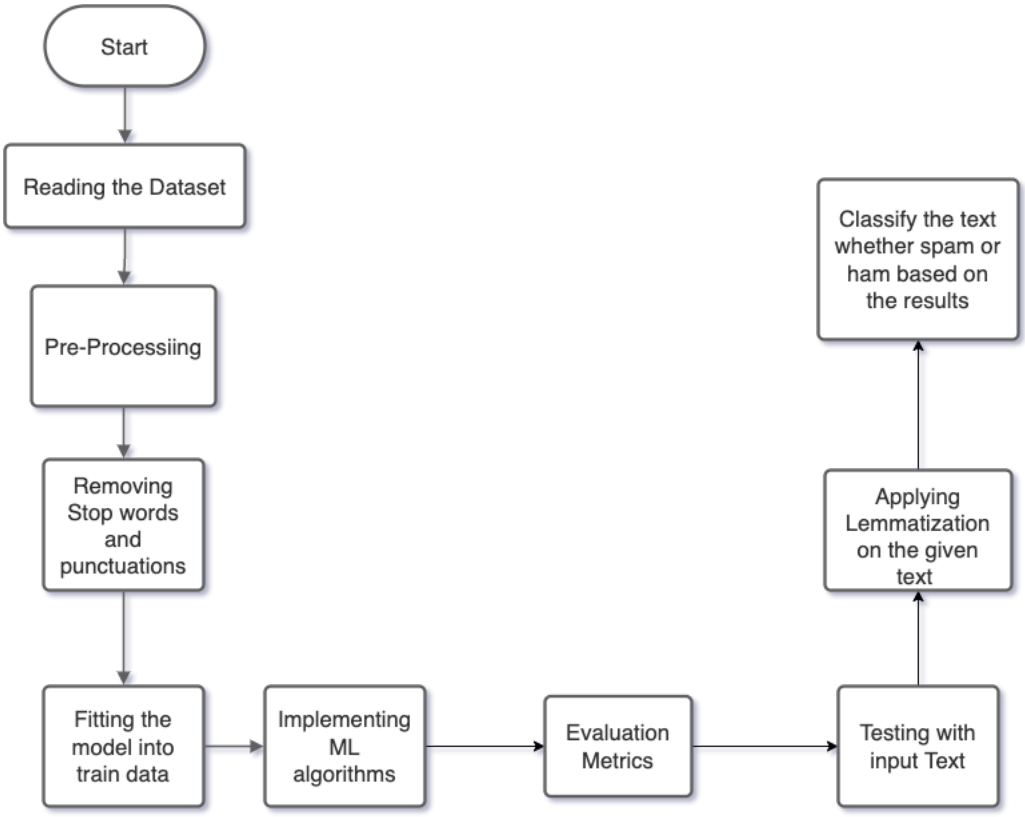
Here, we have collected more than 1000 messages that are identified as ham & spam filled. Then, we will clean the data by eliminating punctuations and stop words using NLTK library.

We will use lemmatization i.e., normalization technique on created tokens to retrieve root words. In addition, we will apply count vectorization to eliminate words that are infrequent in the data. Furthermore, to predict the text messages we will use naïve Bayer classifier and SVM classifier from python "sklearn" package.

- **Features:**

The key feature of this project is to identify messages as either spam or ham. At first, the less common terms will be eliminated from the data once pre-processing and normalization techniques have been used. In order to determine if a message is spam or ham, we must train the model using the cleaned data and apply ML algorithms. As an example, the model will eliminate all the stop words and unnecessary text when we provide it with a message. Finally, after applying lemmatization to the message, the chosen algorithm will determine if the message is spam-or-ham filled.

Workflow diagram:



References:

Kaggle website:

https://www.kaggle.com/datasets/team-ai/spam-text-message-classification

Journal document:

https://jusst.org/wp-content/uploads/2021/08/Classification-of-Spam-Text-using-SVM.pdf

GitHub:

https://github.com/Naveed945/Spam-Message-classifier