SMS Spam Classification using Machine Learning

Project Report

Soumyadeep Poddar

Introduction:

In this project, we explore the detection of SMS spam using a logistic regression approach. We analyze a SMS spam collection dataset and employ various text processing techniques to extract relevant features, such as character count, presence of numbers, and top spam words. Leveraging the Generalized Linear Model (GLM), we build a classification model to distinguish between spam and ham messages. The model's performance is evaluated through confusion matrices and AUC-ROC scores, providing insights into its ability to classify messages accurately. Additionally, we visualize the odds ratios of model features to understand their impact on predicting spam.

Objective:

The main objective of this project is to develop an effective SMS spam detection model using logistic regression. By preprocessing SMS messages and extracting relevant features such as character count, presence of numbers, and top spam words, we aim to build a model that accurately classifies messages as spam or ham. Through model evaluation using confusion matrices and AUC-ROC scores, we aim to assess the model's performance on both training and test data. Furthermore, we seek to visualize the odds ratios of model features to uncover the significance of each predictor in identifying spam messages.

Methodology:

1) Data Collection & Loading:
   - ✓ The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according being ham (legitimate) or spam.
   - ✓ The files contain one message per line. Each line is composed by two columns: v1 contains the label (ham or spam) and v2 contains the raw text. The dataset can be accessed here Kaggle.
2) Splitting into training and test sets:
   - ✓ Splitting the dataset into training and test set. The training set is used to train the logistic regression model. After training the model, it must be evaluated on unseen data, which is where the test set, containing previously unencountered data, becomes crucial.
3) Feature Engineering:
   - ✓ Incorporated into our project, we will employ fundamental text mining tools like character counting and regular expressions to extract deterministic elements from the textual content. These extracted features will subsequently play a role in our classification task.
4) Filtering out words and text analysis:
   - ✓ We performed essential text processing tasks on the SMS messages dataset. We began by tokenizing the SMS messages, a process that involves breaking down the messages into individual words or tokens.

- ✓ We also removed common stopwords. These stopwords, being frequently occurring words with limited semantic value, were filtered out to enhance the quality of our analysis.
- ✓ Moreover, we implemented a step to filter out words with a character length of less than 3. This ensured the removal of very short words that might not convey substantial meaning in the context of our analysis.
- ✓ Subsequently, we computed and examined the frequency of each word separately for both ham and spam messages.

5) Model building and evaluation:
   - ✓ We performed Simple Logistic Regression to model the probability of an SMS being SPAM.
   - ✓ We see that logistic regression performs very good on training as well as test data.

Conclusion:

In conclusion, this project presented a comprehensive approach to SMS spam detection using logistic regression. By utilizing the Logistic Regression model, we successfully built a predictive model capable of distinguishing between spam and ham messages.

Limitations:

It is important to note that this analysis is based on a specific dataset. While logistic regression proves to be an effective technique for SMS spam detection, its performance might be influenced by the complexity of messages. Some sophisticated spam messages might not be fully captured by the selected features, leading to potential misclassifications. Additionally, the success of the model relies heavily on the quality of the training data, and any bias or imbalances present in the dataset can impact the model. Future iterations of this project could explore more advanced techniques, such as natural language processing (NLP) and machine learning algorithms, to address these limitations and further enhance the accuracy and robustness of the SMS spam detection system