207 Project Proposal

SMS Text Classification

Section 2 : Prof. Tanya Roosta

Members:
Hillary Chang, hillary_chang@berkeley.edu
Raji Chandrasekaran, raji.chandrasekaran@berkeley.edu
Song Gao, song_gao@berkeley.edu

Motivation:

Our project focuses on the question: How effectively can machine learning models classify SMS messages as spam or legitimate? The goal is to build a model that can be used to automatically identify unwanted or malicious messages based on their text content.

This question is both interesting and impactful because SMS spam remains a widespread issue, often used for scams, phishing, or unwanted advertising. Unlike emails, SMS poses unique challenges due to shorter message length, informal language, abbreviations, and lack of contextual metadata which makes classification more challenging. Solving this problem effectively will improve user experience and security and demonstrates the practical value of machine learning in handling real-world text data with noise and ambiguity.

Data:

We found a public dataset called SMS Spam Collection which contains 5,574 text messages, each labeled as either "ham" or "spam", classifying the messages as legitimate or not. 425 SMS spam texts came from Grumbletext forum, a UK forum where cell phone users make claims/report spam messages they have received. 3,375 of the ham messages came from the NUS SMS Corpus, a dataset collected by the National University of Singapore, which mainly contains messages from Singaporeans or mostly students attending the University. 450 of the ham messages were taken from Caroline Tagg's PhD thesis. 1,002 SMS ham messages and 322 spam messages came from SMS Spam Corpus v.0.1 Big. The dataset is text-based and each line is formatted so it contains a label, then the actual message. The dataset has no missing values, and it is mainly used for classification or clustering. tasks in computer science and natural language processing. The texts contain normal conversations to scam content.

Source Link: https://archive.ics.uci.edu/dataset/228/sms+spam+collection

Related work:
A recent study titled "SMS Spam Detection and Classification to Combat Abuse in Telephone Networks Using Natural Language Processing" (Oyeyemi & Ojo, 2024) uses BERT-based feature extraction combined with machine learning models like Naive Bayes and SVM to classify SMS messages. Their approach achieved a high accuracy of 97.31%, with Naive Bayes performing best in both speed and accuracy. This work demonstrates the effectiveness of combining modern NLP techniques with traditional ML models for spam detection.
Reference: https://arxiv.org/pdf/2406.06578

Methodology
The dataset is ideal to run classification algorithms. The dataset is provided as a prelabeled dataset that helps us with validation. The common classifications algorithms that we are planning to try for are:

- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machines

We will also apply a combination of algorithms to attain a better accuracy score. Evaluation techniques such as K fold and stratified K fold to better fit the model to the dataset.

Project GitHub repo: https://github.com/hillarychang/datasci207-proj