

Capstone project proposal

Finalize one Capstone idea based on the feedback you got from your mentor and peers on your Section 1 submission, and on your newly acquired understanding of the tools and data wrangling. Submit a project proposal - a short (1-2 page) document that answers the following questions:

1. What is the problem you want to solve?
2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?
3. What data are you going to use for this? How will you acquire this data?
4. In brief, outline your approach to solving this problem (knowing that this might change later).
5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

1. The growth of mobile phone users has led to a dramatic increase in SMS spam messages. Therefore, there is a need for reliable filtering techniques capable of detecting spam messages from ham messages. The objective of the current capstone project is to develop an automatic filtering system using natural language processing (NLP) and machine learning techniques for SMS spam detection.

2. Potential clients for an SMS spam detection system are mobile phone service providers. Mobile service providers can use this system to automatically filter out spam messages at the first place so that their clients do not receive those spam messages which potentially leads to more customer satisfaction. Using this filtering system, the mobile service providers can trace the sources of spam messages and implement strategies to eliminate/reduce the rate of spam messages which are generated/received.

3. The dataset for this project originates from the UCI Machine Learning Repository which is publicly available. 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.

4. The initial steps toward development of an SMS spam detection framework is to turn the text messages into numerical values/features that could serve as the input for machine learning models. Using NLP techniques, it is feasible to generate a dictionary by detecting the most common words in the dataset to turn each raw text message into a vector showing the presence/absence of certain words. This is also important to give weights to the words based on their level of importance (e.g. occurrence in spam/ham texts). It should be noted that stop words have to be detected and removed from the texts before performing the analyses. Following vectorization of texts, the numerical representation of the data could be input to train machine learning models. In this study we explore the application of various machine learning models (K nearest neighbor, random forest, AdaBoost, decision tree, neural networks, etc) and evaluate them based on their performance using various metrics (accuracy, F1 score, precision, etc).

5. The capstone project final report will consist of final codes, exploratory data analysis of data set, trained models/classifiers and their performance metrics.