**Anomaly Detection in SMS Applications**

Robert Chavez, Kiera Conway

**CSC 723 – Machine Learning for Cyber Security – Project Topic Proposal – Spring 2023**

**Proposal Content**

Our group’s final project for the Spring 2023 semester of CSC 723 will focus on machine learning models that are geared towards anomaly detection. The chosen field of study for our machine learning models is Short Message Service (SMS) spam. We will be leveraging various SMS spam datasets to train, test, and analyze various classification algorithms to find the most precise option to help determine if an SMS is considered spam. SMS has become one of the most popular sources of medium used for marketing and advertising. We feel that it is important to develop an understanding on how one can utilize machine learning that can accurately detect SMS spam. The precision and accuracy will be determined using a combination of model scores, confusion matrices, and other applicable model metrics. While the math deriving most of these metrics will be computed using the Python library , any formulas used in the project will be thoroughly explained for reader clarity. Furthermore, the training and machine learning algorithms will be created using the Python scripting language and submitted alongside the final project report.

All scripts written for this project will be created using the Jupyter Notebook environment Kaggle. Our group has determined the chosen format to be submitted will be in the “.ipynb” the Kaggle file format. Our group selected Kaggle as our main environment due to the abundance of datasets, the ability to directly import said datasets, and the ease of facilitation between group members. While each algorithm will be trained and tested in a separate notebook, the final report will consolidate the data in an organized manner. Any additional materials which require group collaboration such as outlines, notes, and resources. In an effort to better assist our team with coordinated efforts we will be leveraging a shared GitHub repository in which all materials, notes, and project data will reside for the project remainder.

Since we will be classifying discrete, fixed and binary values such as SPAM vs HAM. Our group will use various Supervised Classification algorithms. The three algorithms chosen to test, train, analyze, and compare our results are K-Nearest Neighbors (KNN), Decision Tree, and Naive Bayes (including Bayes’ Theorem to calculate the joint density of dependent events). While each algorithm script may contain additional unique libraries the primary libraries utilized in our project will be , , and . Our group will ensure each algorithm will be trained and tested using the data set SMSCollection.csv which is derived from the spam-or-ham dataset from Kaggle. Prior to testing, the dataset will be thoroughly reviewed in Python by the group. The purpose of this review will be an initial effort to obtain general information related to the data set such as specific observations and statistical information relevant to the dataset. Post review, the team will ensure each step during the training and testing process with be explained using code examples for support and documentation purposes. Lastly, all three algorithms will be tested to compare and contrast which is the best suitable option for the SPAM classification algorithm.

All team member contributions will be equal as we share responsibility for project activities not limited to research, scripting, consolidation, writing, and reviewing. To ensure team success and accountability the team set up a tentative schedule to track progress, milestones, and weekly meeting notes via a scheduled Discord call. The team will leverage these weekly meeting sessions to discuss plans and strategies relevant to the project. All research materials are to be submitted to a shared ‘Data Dump’ folder. In addition, script work is done via a shared notebook and written documents are regularly updated on our shared GitHub repository. Furthermore, the algorithms and report sections will be split evenly to maintain workload symmetry. As previously mentioned, all tools have been selected with cooperation as the driving factor. In conclusion, with our project findings and research we hope to help work towards minimizing spam advertising and marketing SMS spam by leveraging today’s powerful tools and algorithms.

**Resources**

<https://www.kaggle.com/datasets/arunasivapragasam/spam-or-ham>

<https://dl.acm.org/>

<https://www.sciencedirect.com/>

<https://scholar.google.com>

<https://ieeexplore.ieee.org/>