

**Rochester Institute of Technology**



**CSEC 520/620 Cyber Analytics and Machine Learning**

**Project 2: Detecting SMS Spam**

**Description:**

The goal of this project is to learn how to process unstructured text data that we encounter in many real-world cyber analytics problems. Your main tasks are to convert text to numerical vectors and create a classifier to detect an SMS message if it is a Spam or not.

**Steps:**

1. Read the readme file about the data, read the research paper on the dataset
2. Using NLTK or similar NLP libraries to process each message, and create a term-document matrix ***M*** based on tf-idf values.
3. With the row of ***M*** fixed, each column is a document vector, i.e., a data point like those in the first project.
4. Apply the *k*-means algorithm you implemented in the first project to this data set.
5. Implement the *k* nearest neighbor algorithm and apply it to this data set. Compare your results with the results reported in the research paper.

**Deliverables:**

1. All source code of your implementations
2. A description of how to compile and/or run your program including program dependencies
3. A discussion on the performance of your implementations in Step 4 and 5