# Foundations of Machine Learning Midterm Report: Paraphrase Identification

Christine Cho

October 27th, 2022

## 1 Introduction

The goal of this project was to train a LR or SVM model for the task of classifying whether or not two sentences were paraphrases of each other. I worked on processing the data given, feature extraction, finding which model was best for accuracy (SVM), and tuning the model to improve performance. All of this was done in Google's Colab environment.

## 2 Features

The features and methods of feature extraction were taken from different papers and GitHub repositories regarding paraphrase identification. These features include total number of words in a sentence, the sentence length, the absolute value of the difference between sentence lengths, the longest common sub-sequence, the number of unique words shared by the two sentences, the total number of words from both sentences combined, the ratio of unique words to the total number of words, the minimum edit distance, the euclidean distance, cosine similarity, and different containment values using different n-grams.

## 3 Data Preprocessing

The test, train, and dev text files that were provided were read in and converted into DataFrames using the pandas library. The sentences in each row were stripped of punctuation and white-space at the front and end. Functions were created to extract features and store them in their own DataFrames. Most of the features extracted were used except for some containment values of different n-grams. The features were also standardized for scaling.

## 4 Algorithms and Libraries

Algorithms to calculate the features like cosine similarity, euclidean distance, etc, were provided by libraries like sklearn and scipy. The machine learning models (SVM and logistic regression) were constructed and trained using sklearn as well. Other algorithms like finding the containment value were implemented by hand and found from other papers and Github repositories.

## 5 Results

With this project I gained more experience in using the sklearn and pandas library. The initial process of reading through papers and looking at past projects for paraphrase identification was not a new experience, but I got to learn about new concepts and algorithms used in developing NLP tasks. The limitation on using logistic regression and/or SVM algorithm allowed me to develop a better understanding of two.