CIS600 Principles of Social Media and Data Mining
Spring 2025

Assignment 3 – Twitter Sentiment Analysis

Due: 23:59pm on Wednesday, April 2, 2025
Total Marks: 10

ALL submissions should be via Blackboard. No hand delivery will be accepted. Late submissions will be graded out of 50%. This assignment must be done individually.

What to submit:
Your code in the form of ipython notebook file should be submitted. You will also submit your report as a PDF.

Objectives:
This assignment will provide you hands-on experience with forming social media text representations and the use of text categorization for sentiment analysis.

The three datasets provide experience with different types of social media content. This assignment also gives you practice with a type of question that you will see on the final exam.

This assignment requires you to perform the following tasks:

1. [30%] Data preprocessing and normalization (lowercase transformation, punctuation and special character removal, stop words elimination, short and long word deletion, stemming, lemmatization, removal of emojis, emoticons, URLs, HTML tags etc..)
2. [20%] Vectorization (you may use unigrams, bigrams or POS tag)
3. [20%] Feature selection or feature transformation
4. [10%] Classification model building (training and testing)
5. [10%] Model results and performance
6. [10%]Write a report that discusses your findings and experience with the text classification approach to sentiment analysis

Dataset
For this assignment, we will use tweets about Apple corporation that were extracted from a Twitter dataset created by Sanders Analytics. The file **twitter-sanders-apple2.zip** contains 479 tweets in a csv format. It has two categories: Pos (163 tweets that express a positive or favorable sentiment) and Neg (316 tweets that express a negative or unfavorable sentiment).

Experiments

In this assignment, you may carry out two experiments as follows:

Experiment #1:

> Create baseline representations for twitter-sanders-apple2 datasets. The representations are defined as follows: unigrams, binary features, threshold=3.
>
> Test your baseline representation for each category (Pos, Neg) using the following classifiers: Naïve Bayes and SVM with the linear kernel. Report Precision, Recall, and F1 for each category (e.g., Pos and Neg) in baseline representation.

Experiment #2: POS features

> Second, create a new representation using words tagged with their parts of speech (POS tags).
>
> Compare and discuss the results of experiment #1 and experiment #2.