# Lab + 2
# Sentiment Analysis of Microblog Data Streams

**Due date:** 23 Oct 2020

**Presentation Date:** 23 Oct 2020

## Objectives:

The sentiment of tweets is another important task that helps to determine the users' perceptions of organizations, events or products. This assignment aims to develop a classifier to classify input stream of tweets into multiple sentiment classes (the simplest being just the positive and negative classes). One important resource in sentiment analysis is the dictionary of sentiment vocabulary, or sentiment lexicon, which may vary according to time (new sentiment terms may emerge) and classes (same term may have different sentiments with respect to different classes). In addition, it should also explore temporal information in determining the sentiment of incoming tweets.

## What You Need to Do:

The assignment will implement a module to perform sentiment analysis of incoming microblogs streams. It should incorporate the following functions:

1. **Basic Sentiment Classifier:**
   - It will process the training dataset to train the classifiers using any suitable machine learning technique.
   - The initial classifier will be trained based on text features only using a basic set of sentiment lexicon.
   - The basic classifier will assign a new input tweet into 3 classes of: positive (+1), negative (-1) and neutral (0).
   - The ability to predict organization (e.g. Apple) given a Tweet
   - The ability to perform sentiment analysis with respect to given organization classes.
2. **Enhanced Functions**:
   - Able to explore temporal information in determining the sentiment of incoming tweets.
   - Expand to more than 3 sentiment classes (say in 5-point scale)

## What You are Given:

- You are given a set of tweets for 4 organizations: Apple, Google, Microsoft and Twitter. All the tweets were published within the period of 15-20 October 2011. The number of tweets for each organization is shown in the following table.

| Training Set: | Testing Set: |
|---|---|
| Apple : 981 | Apple : 109 |
| Google: 788 | Google: 88 |
| Microsoft: 778 | Microsoft: 86 |
| Twitter: 866 | Twitter: 96 |

**Presentation and Online Testing:**

- You will need to present your work within a 20-min session, including question answering, during which you will present your work using ppt and demonstrate the effectiveness of your software on your Notebook.
- Test tweets will be used to test the performance of your classifier during online evaluation.

**Report:**

You need to submit before the deadline:

1) A report of not more than 8-pages. It should include the program structure, details of your classifier, training and testing procedures. You also need to include details of your testing with tabulated results showing the effectiveness of your classifier.
2) A short ppt file (for not more than 8 mins of presentation) that includes sufficient details for the instructors to understand the details of your program and testing.
3) Source codes of your implementation (on GitHub). There should be a readme.txt file to describe how to run your program. Your program should be able to take in several new tweets online from the interface and return the classification results.

**Remarks:** (a) Techniques, flexibility and effectiveness of system is most important; UI is less important and hence do not spend too much effort on refining your UI. (b) All members are required to present some aspects of the system when asked during the project presentation and demo. (c) Extra marks will be given for excellent assignments.

**Consultation:**

Any questions regarding this assignment, please consult:

- Mr. Yang Qi

** **Late Submission Policy**: We impose the following penalties for late submissions. (a) Late but within 24 hours: 25% reduction in grades. (b) Later but within 3 days: 50% reduction in grades. (c) After 3 days: zero marks.