**King Fahd University of Petroleum & Minerals**

College of Computer Sciences and Engineering

Information and Computer Science Department

**ICS 490: Special Topics in Computer Science**

Term Project

[Twitter Sentiment Analysis]

**Posted:** Tuesday 21st November 2017  
**Deadline:** Thursday 14th December 2017 @ 11:59 PM (Before Midnight)

**Abstract:** As social media is maturing and growing, **sentiment analysis** of online communication has become a new way to gauge public opinions of events and actions in the world. Twitter is a main source for many people to “voice” their ideas in extremely short messages!

They can choose to "retweet" or share a tweet, to promote ideas that they find favorable and elect to follow others whose opinion that they value.

**Initial Dataset:** Hand-curated **Twitter** sentiment dataset published by **Sander’s Lab**. It contains tweets from 2007 to 2011 that mention one of four major Tech companies. Sander’s Lab manually

assigned labels for each tweet as either “**Positive**”, “**Negative**”, “**Neutral**”, or “**Irrelevant**”[[1]](#footnote-1). “**Positive**” and “**Negative**” indicated whether or not the tweet showed a generally favorable or unfavorable opinion toward the mentioned company. A “**Neutral**” labelling indicated that the tweet was either purely informative or the opinion of the tweet was otherwise ambiguous. An “**Irrelevant**” labelling indicated that the tweet could not be determined to fit into any of the other classes. This often indicated that the tweet was not written in English or that it was clearly spam.

**Tasks:**

1. Investigate Sander’s sentiment corpus and study sentiment toward technology companies in **Twitter**. Implement **3 different** multi-class classifiers to classify this data.
   1. Investigate the possible features to use to train your 3 classifiers.
   2. Provide a summary of your evaluation strategy in terms of training, validation and testing for your 3 classifiers.
2. Build your own twitter dataset for sentiment analysis in Arabic language where labels should be: “**Positive**”, “**Negative**” and “**Neutral**”.
   1. Make sure to have at least **5000 tweets** in your database
   2. Make sure to label all the collected tweets.
3. Implement a classifier to classify your generated data. Make sure to use an ensemble classifier that will consist of:
   1. Support vector machines.
   2. Random forests.
   3. Multinomial Naïve Bayes.
   4. Softmax classifier
4. Write one report with analysis on the results obtained for each dataset. So, you will need to submit two different reports.

1. N. J. Sanders, "Twitter Sentiment Corpus," Sanders Analytics (Available online at: http://www.sananalytics.com/lab/twitter-sentiment/) [↑](#footnote-ref-1)