**F21AA Applied Text Analytics: Coursework 1**

**Handed Out:** Wednesday 15th January 2023

**Work organisation:** group work,in groups of 3-4 students.

**What must be submitted:** Well documents Python code

**Submission deadline: 11:55 pm Friday 10th March 2023 -- via Canvas/ group space**

**Worth**: 20% of the marks for the module**.**

**Objectives and Problem Formulation:**

Studies have shown that sentiments from Twitter could be very useful for businesses and government entities. For example, through analysis of tweets, one could predict the early onset of a disease in a region, stock market change, results of elections, sales projection and customer complaints.

In CW1 you are invited to apply essential text analytics tools to the problem of ‘Twitter Sentiment Analysis’. In particular, you are required to analyse tweets related to ‘**Technology and Innovation in UAE’[[1]](#footnote-1).**

The main purpose is to provide insights on sentiments and main topics discussed from the collected data. You are also required to build a categorization system that can automatically assign sentiments to incoming unlabeled tweets.

*In this coursework you will gain experience in:*

* Creating a developer account with Twitter to access the Twitter API
* Using Tweepy, TextBlob and various other built-in python libraries to extract, preprocess and analyze tweets.
* Collecting your data set from tweets to train/test your models.
* Applying visualization and analytics tools to gain insights from tweets and to categories any new tweets based on their polarity.

**Implementation and Requirements:**

You should use Python (and any additional libraries) to conduct the following steps:

1. **Data Collection: (15%)**
2. Create your own dataset by extracting tweets from twitter using the [Twitter API](https://developer.twitter.com/en/support/twitter-api). To do this you will first need to create a [developer account](https://developer.twitter.com/en/support/twitter-api/developer-account#faq-developer-account) with twitter[[2]](#footnote-2).
3. Once the developer account is approved, you can use [Tweepy](https://docs.tweepy.org/en/latest/), the built-in python library to regularly extract the tweets and other relevant metadata required to build your own dataset.
4. There is no set limit on the number of tweets you are required to collect[[3]](#footnote-3). You should at least collect 100 tweets related to your topic.
5. Pick hashtags and search terms that give you the most relevant tweets for your problem formulation.
6. **Data Labelling: (15%)**
7. Choose relevant sentiments that best represent tweets in your dataset and then label individual tweets accordingly. For example, you could decide to hand-tag the tweets as positive, negative and neural. Split your data into 80% for training and 20% for testing. Other alternatives for labeling are also possible (using an existing labeled data set, using sentiment lexicon,…. ). Please justify and outline your choices.
8. You could decide to use TextBlob[[4]](#footnote-4) or Vader[[5]](#footnote-5) to perform an initial analysis of tweet polarity and provide the top 5- 10 tweets representing each of these polarities (The different polarities are commonly denoted by +1: positive, -1: negative and 0: neutral). Document your findings and make conclusions for your next steps.
9. **Text Analytics pipeline on tweets: (20%)- free style**

Apply a text analytics procedure of your choice. Always outline and justify your methodology.

1. Create a text analytics pipeline for text processing, representation, classification and evaluation.
2. Text processing techniques could include: tokenization, stemming, normalization and stop-word/punctuation removal, etc.
3. Representation and feature extraction can include using n-gram features vs. unigram features; Vector space representation (binary, frequency count & tf-idf).
4. Compare the performance of the classifiers based on precision /recall.
5. **Visualization and Insights from tweets: (20%)**
6. Examine your most prevalent sentiments separately (example positive/negative sentiments). Can you infer any particular observations regarding the topics discussed in the different tweets representing these sentiments?
7. Visualize your results from the previous step using word clouds, pyLDAvis and any other visualization techniques. Document any other observations you have gained with this analysis.
8. **Discussion and conclusion from experiments: (20%)**

Highlight the main findings from steps A-D. Justify important points like: How did you validate the initial labels assigned to tweets in the dataset? Summarize the insights you gained from your experiments about how people feel about **‘Technology in UAE’**. Argue how useful such findings could be in practice.

1. **Research question (PG only):** (One A4 page font Arial 11pt + relevant references) (10%)

Summarize one paper that you have found interesting for your study. The paper should be peer reviewed from a reputable journal or conference. The paper can tackle any of the topic of relevance to your experiments: sentiment analysis, text classification, text labelling, text representations, unsupervised learning,…etc

**What to Submit**

***Upload the following to your group space:***

1. Python notebook that includes all relevant experiments. ***Additionally, please upload a link to a google collab version of the notebook***. Your notebook should be well documented and should have a clear description of the CW tasks. Include also any extra evidence of your work: data sets, scripts, tables, screenshots, etc. or supply a link.
2. Task distribution per group member. Tasks can include implementing different part of the CW, project management, researching academic references and tools, generating visualizations and graphs, report writing and preparing presentation/recording.
3. Prepare a 10min demo of your methodology for steps A-D and discuss your findings (step E). Include graphs and comparison tables when possible. All group members should prepare to answer questions. CW interviews will take place March 14th during lab time and the exact schedule will be announced one week in advance.
4. PG students submit a 1-page report for their research question. (Submit through Canvas link)

Include appropriate research articles and other references to support the academic arguments for the approaches taken in the completion of this coursework.

**Additional Remarks**

* This is a group assignment. Group members will collaborate on data collection, methodology, code, report and research to produce their own work. University plagiarism rules will apply (see below).
* You will find many practitioners resources that can help you in this coursework. Make sure to reference your sources and potentially outline how your approach is different/similar.
* Higher marks will be awarded to novel ideas and original thinking. You are also required to show what you have learned in the course in addition to your independent research and findings.

**Marking**: See detailed marking Rubric on Canvas. Maximum points possible: 100.

**Plagiarism and Collusion**

Students must never give hard or soft copies of their coursework reports or code to students in another group. Students must always refuse any request from another student not in their group for a copy of their report and/or code. It is expected that all group members will have read and write access to the report and code for their group.

Sharing a coursework report and/or code with another group is collusion, and if detected, this will be reported to the School's Discipline Committee. If found guilty of collusion, the penalty could involve voiding the course.

Readings, web sources and any other material that you use from sources other than lecture material must be appropriately acknowledged and referenced. Plagiarism in any part of your report will result in referral to the disciplinary committee, which may lead to you losing all marks for this coursework and may have further implications on your degree. <https://www.hw.ac.uk/students/studies/examinations/plagiarism.htm>

**Lateness penalties**

Standard university rules and penalties for late coursework submission will apply to all coursework submissions. See the student handbook.

**Feedback:** You will receive your marks and detailed feedback during interview time and marks will be posted on Canvas after 14 working days of your submission deadline.

1. Alternatively, you may choose search terms/keywords or hashtag in relation to Technology like ‘Meta Verse’, ‘Innovation’, ‘Sustainability’, ‘Technology Footprint’, ‘Artificial Intelligence’ , ‘Ethics’, ‘Inflation’, ‘COP 28’ , ‘ Museum of the Future’, ‘Bitcoins’, ‘Cyber Attacks’, etc. [↑](#footnote-ref-1)
2. <https://developer.twitter.com/en/support/twitter-api> [↑](#footnote-ref-2)
3. More tweets are better to build models that don’t overfit [↑](#footnote-ref-3)
4. **TextBlob** is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. [↑](#footnote-ref-4)
5. **Vader** is an NLTK module that provides sentiment scores based on the words used. It is a rule-based sentiment analyzer in which the terms are generally labeled as per their semantic orientation as either positive or negative [↑](#footnote-ref-5)