Course Work – MSc Students

COMPSCI5078

Social Media Emotion Data Set

Total Marks – 100 marks & Weightage – 20%
Course work deadline – Monday 9th March 2020 4:30PM

INTRODUCTION

Ps1: This course work is for the students who would like to do an individual project instead of a group project. Do not do both (group and individual)!!

Ps 2: Please note a small fee (less than £10) will be needed to do crowd sourcing.

The objective of this course work is to develop an emotion annotated data set from Twitter and conduct analytics. The objective is to use this data for developing emotion detection systems. So it is important that the data should be cleanly labelled into respective classes.

We recommend students to use Python programming language and also MongoDB for data storage. It is very important that students provide working version of the software, as we need to validate them.

Students submit their **sample data**, **code and report** on or before the specified deadline.

Code and sample data should be provided on the github. Report submission is through the Moodle page for the Web Science course.

The coursework will be marked out of 100. Course work will have 20% weight of the final marks. As the usual practice across the school, numerical marks will be appropriately converted into bands. Final written exam will have 80% weightage, which will be in April/May 2020.

There will be a lecture on ***sentiment & emotions*** for more details (lecture 7). Similarly on ***crowd sourcing*** for other details (lecture 8).

Specific tasks to do

1. [Total 30 marks] Develop a crawler to access a collection of English Twitter data. The idea is to have 150 reasonably clean tweets per class. Six classes with 150 tweetscper class will lead to a collection of 900 tweets

   We will have six classes of data corresponding to

Excitement, Happy (joy, love), Pleasant (+ve feeling), Surprise (sad, frustration, -ve feelings), fear (disgust, depression), Angry. You use hashtags to identify representative classes of data (for example #happy).

a. [10 marks] Use Twitter API for collecting data for emotion labelling. Minimum 900 tweets with at least cleanly labelled 150 tweets per class.

   Organise these into separate files for each class . This will lead to the creation of an emotional data set for developing models

   Provide a table summarising data statistics for each class.

b. [20 marks] Inevitably lot of tweets will be ambiguous and might overlap between the six classes. Need to find a way to avoid this by processing hashtags.
   i. Discuss strategies/rules to collect 'reasonably' clean data. (10 marks) Describe the hashtags or other methods you used.

   ii. In addition to hashtags emoticons (e.g., ☺) could be used. (10 marks)

2. [30 marks] Processing of data. Our objective is to have a clean text (means unambiguous emotion category) so that it can be used for training a classifier for detecting emotions. So each tweet need to be pre-processed.
   i. E.g., Looooove can be shortend to love! (10 marks fro developing methods to clean text)
   ii. Other linguistic anomalies should be removed (10 marks)
   b. [10 marks] Appropriate storage of data, structure and other details. Removal of duplicates. (10 marks)

3. [30 marks] Crowdsourcing
   a. Take a sample of data (e.g., say a random 20 per class) and get crowd sourced annotation. The idea of crowdsourcing is to verify whether our data is clean enough or not!
      i. Crowd sourcing scheme – 15 marks
      ii. Discussion of results – 15 marks

4. [10 marks] Report. Organisation, structuring, language etc.

Report structure

Report should be organised the following way:

0: Front page
Title –
Student Name & Matriculation Number
Source code – github information

Data – github information for code and sample data

1. Section 1: Introduction
    a. Describe the software developed with appropriate details; if you have used code from elsewhere please specify it
    b. Specify the time and duration of data collected

2. Section 2: Data crawl & Rules
    a. Use Twitter API for collecting data
        i. Specify the APIs used
            1. Please do not include entire code here; just main description of the function
            2. Along with a short description/justification
    b. Aim is to have a minimum 900 tweets with at least 150 tweets per class.
            1. Give data statistics. A table with
                a. Total and individual class distribution along with time period in which data collected.

    c. Processing of Tweets:
        i. Discuss strategies/rules to collect 'reasonably' clean data. A clean tweet is one which is unambiguously belongs to the correct emotion class. The original tweet text should be provided;

           Tweet id, text of tweet, created_at information to be stored

           Inevitably lot of data will be ambiguous, overlap between classes. Need to find a way to avoid this by processing hashtags. Elaborate on the rules and rationale

           Tabular or histogram representation of the effects of rules.

        ii. Rules to use emoticons for labelling data. Explain the rules and specify the number of tweets collected by these rules

           Tabular or histogram representation of the effects of rules.

           Total data with emoticons
           Individual class data
           How many removed from each class.

           Tabular or histogram representation of the effects of rules.

3. Crowdsourcing method
    a. Crowdsourcing details
        i. Data used in crowd sourcing. Interface of crowd sourcing.
        ii. Questions asked and in what order. Provide a rationale

iii. Statistics on agreement.

b. Results and discussion

Submission

Deadline: Monday 9th March 2020 4:30PM

What to submit

1) Report as a pdf file. (Please submit this jus for the report link)
2) A github information on data and code
    a. Software (runnable version, readme info, and also properly commented). It is important that software is runnable with minimum effort for the markers.
    b. Data – provide a sample data. Use a CSV format or MongoDb. Importantly your software should be able to run on this sample data, without much hassle.

Where to submit

1) Moodle
2) Code and sample data on github