University of South Wales
Prifysgol De Cymru

Faculty of Computing, Engineering and Science

Assessment Cover Sheet and Feedback Form 2020-21

Module Code: MS4H05	Module Title: Text Mining and Natural Language Processing	Module Team: Moizzah Asif, Penny Holborn

Assessment Title and Tasks: Coursework and Practical demonstration of code	Assessment No. 1 & 2

Date Set: **21-Jun-21**	Submission Date: **12-Jul-21**	Return Date: **02-Aug-21**

IT IS YOUR RESPONSIBILITY TO KEEP RECORDS OF ALL WORK SUBMITTED

Marking and Assessment
This assignment will be marked out of 100% The two parts (Part 1 & 2) of this assignment contribute to 70% and 30% respectively of the total module marks.
Learning Outcomes to be assessed (as specified in the validated module descriptor https://icis.southwales.ac.uk/): 1) To understand the methods for analysing unstructured data and explain the wider context of their value in Data/AI. 2) To utilise and critically evaluate a variety of NLP algorithms to assess and interpret real-world complex unstructured data.
Provisional mark only: subject to change and / or confirmation by the Assessment Board

MS4H05 Coursework & Demonstration: 2021

Part 1 - Coursework

Submission deadline: Monday 12th July 2021 - 21 00
Contribution to module assessment: 70%

Part 2 - Demonstration

Submission deadline: Monday 12th July 2021 - 21 00
Contribution to module assessment: 30%

Submission Guideline

The assessment is divided into two main Parts: Coursework and Demonstration.

1. **Coursework**

   You are required to:

   (a) submit a report with appropriate programming elements and code in at least one of the following programming notebooks (markdown reports) on BB assessment:

   - **Jupyter Notebook** - .ipynb along with its html or pdf version. Please provide these for the coding and accompanying reporting elements in Python (if any).

     or

   - **R Notebook** - .rmd along with html or pdf version. Please provide these for the coding and accompanying reporting elements in R (if any).

   (b) attempt all 4 tasks (A, B, C and D)

   You may attach appendices with pdf or html files of your report.

   If you choose to perform tasks across programming platforms, then you need submit an extra pdf file which combines notebooks for each task in sequence and can be considered as the final report of your coursework.

   Note: If .rmd, .ipynb or any other file formats can not be uploaded directly on BB assessment item, then please create a zip folder for only those files and upload the folder instead of the incompatible format files.

   You are required to upload at least one file outside of the zip folder which can be read and marked directly on BB. Ideally, this will be the final report in either pdf or html format.

2. **Demonstration** You are required to record a 5 minutes' walk through video of your programming notebook/s from Assignment Part 1 - coursework, and submit the recording by the deadline under

this BB assessment.

Any length of recording over 5 minutes will neither be watched nor marked, so please ensure that your recording stays within the time limit.

You will be invited for a 5 minutes Q&A session by the course team after the submission deadline.

The Q&A will be held on **16-07-2021**.

The invite to the event will be shared with you closer to the date.

Instructions

- You are required to use Panopto recorder to record the video. Panopto provides both the web client as well as desktop application to create recordings.
- The desktop application can be downloaded via uniApps: `https://uniapps.southwales.ac.uk/` or from southwales panopto cloud using the link provided in the next line.
- Please ensure that you have logged in to panopto using USW credentials before recording from desktop app.
- Alternatively you can log on to `southwales.cloud.panopto.eu` and use the web client (create → panopto capture).
- Once you have recorded your notebook walkthrough video, please set the access rights/permission to 'Anyone at your organisation with the link'.
- Copy the shareable web link to your recording and paste it onto a text document which you will submit under this assessment as a separate document.
- Please ensure that you have tested the validity of the link before submitting.

Aim of the assignment

You are required to apply text mining and natural language processing methods, algorithms and techniques to answer the questions in each part of this coursework.

You will be expected to use concepts that have been taught in the module or are relevant to:

- the aims and objectives of this module as outlined in the module descriptor,
- the concepts taught in the module.

You are provided with a dataset so that you may be able to demonstrate these concepts' application and usability using programming code and accompanying report style elements.

Expectations from your reports (notebooks/markdown report)

There is set marks allocation on the readability and coherence of the descriptive component of your report and its integration with the coding aspect of your work.

Similarly, the readability and nature of informative feedback of your code, effective visualisations and relevant outcomes of programming code to supplement the reporting elements have a marks quota.

The **Reporting Task** outlines the word count for written components of your submissions (i.e. 1500). However, computer programme codes can not be measured in word count or lines of code. Hence we offer a flexible approach towards the lines of code and word count for reporting element that you may present in the final report. This figure is just provided as a reference to help you find a balance between code and written element of your report.

You should be mindful that your code is coherent and achieves the desired tasks. Instead of screen length, number of lines of code or word count, your code programme/s and script/s will be mainly measured against strength of analysis and meaningful outputs.

Please ensure that: any adaption and re-purposing of code that is not yours; and import of non-standard/non-native programming language packages and libraries that are not yours, are duly cited and referenced.

Plagiarism and any other type of academic misconduct is taken very seriously at USW, hence you are advised to familiarise yourself with good academic practices. You may find useful resources on USW advice on good academic practices.

Data Description

Introduction and Background

The dataset provided with this assignment is called 'CMU-MisCOV19' and it comes from a research project at Center for Machine Learning and Health at Carnegie Mellon University. The work was presented under the following title at CIKM 2020[1], 'Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset'[2]. Part of the abstract from this paper is presented as follows to build your understanding as to why was this data collected and annotated.[3]

> "*From conspiracy theories to fake cures and fake treatments, COVID-19 has become a hotbed for the spread of misinformation online. It is more important than ever to identify methods to debunk and correct false information online. In this paper, we present a methodology and analyses to characterize the two competing COVID-19 misinformation communities online: (i) misinformed users or users who are actively posting misinformation, and (ii) informed users or users who are actively spreading true information, or calling out misinformation. The goals of this study are twofold: (i) collecting a diverse set of annotated COVID-19 Twitter dataset that can be used by the research community to conduct meaningful analysis; and (ii) characterizing the two target communities in terms of their network structure, linguistic patterns, and their membership in other communities.*"

Description

To create this dataset, authors used a diverse set of keywords to filter tweets through Twitter search API. For the annotation process, 17 categories were identified and the tweets were annotated manually. A codebook on annotations and categories, created by the authors, has been provided along with this coursework brief. Please refer to this codebook to familiarise yourself with the categories. The list of categories that these tweets have been categorised/annotated as is provided below:

1. Irrelevant
2. Conspiracy
3. True Treatment
4. True Prevention
5. Fake Cure
6. Fake Treatment
7. False Fact or Prevention
8. Correction/Calling out
9. Sarcasm/Satire

[1] https://www.cikm2020.org/

[2] Memon, Shahan Ali, and Kathleen M. Carley. "Characterizing covid-19 misinformation communities using a novel twitter dataset." arXiv preprint arXiv:2008.00791 (2020).

[3] An open access copy of this paper has been provided with the assignment brief.

10. True Public Health Response
11. False Public Health Response
12. Politics
13. Ambiguous/Difficult to Classify
14. Commercial Activity or Promotion
15. Emergency Response
16. News
17. Panic Buying

The study mentions that 4573 tweets were annotated, and the annotations were made publicly available. However, at the time of data extraction for this assignments, some of these tweets or their authors' accounts had either been suspended or taken down by Twitter, or the privacy settings had changed. Therefore, the number of tweets provided for this assignment is slightly less than those annotated in the study.

Data privacy

Since the data consists of tweets' text, the authors (Memon and Carley), while adhering to Twitter's terms and conditions did not provide full tweet JSONs. Instead they made the tweet IDs and annotations publicly available at the following URL: `http://doi.org/10.5281/zenodo.4024154`.

The tweets text provided here in the assignment has been specifically downloaded from Twitter's search API for this assignment. Please be mindful that the data provided here can not be shared publicly or for any other purpose other than this assignment.

You should make every effort to ensure that you do not post or refer to these tweets' text outside this assignment and display the entire text of a tweet only if and when required with in this assignment.
If you have any questions around data privacy and handling then please free to point them to the module leader.

Assignment Part 1 - Coursework

Task A - Text mining and pre-processing 10 marks

Use appropriate text pre-processing methods to prepare the data to perform exploratory analysis as required in the next task.

The pre-processing steps should include the following:

1. replacing user mentions (i.e, @⟨twitterhandle⟩) with '@user'
2. removal of stop words
3. normalisation of text
4. remove the hashtags which were used to filter the tweets' text. These hashtags are listed below:
   (a) #nCoV20199,
   (b) #CoronaOutbreak,
   (c) #CoronaVirus,
   (d) #CoronavirusCoverup,
   (e) #CoronavirusOutbreak,
   (f) #COVID19,
   (g) #Coronavirus,
   (h) #WuhanCoronavirus,
   (i) #coronaviris
   (j) #Wuhan

   please note you should not delete them entirely from the data, as these might be useful for the next tasks.

You should not hesitate from including any other text pre-processing steps that you may deem appropriate to supplement your approach in performing next tasks.

Task B - Text Analysis 20 marks

Perform necessary text analysis to answer the following questions.

1. What is the distribution of word types in each category and the entire dataset?
   You may want to mention any assumptions that you would use regarding word types, before describing the distribution. These can be based on how you have pre-processed the text in previous task.
2. Are there any similarities among each category's vocabulary?
3. What is the over all distribution of nouns and noun phrases, and verbs and verb phrases, in the entire corpus?
4. Which annotated category has the highest number of user mentions?
5. Which annotated category uses the most hashtags in a tweet on average?
6. What are the top 3 hashtags for each category, including and excluding the hashtags used to filter the tweets?

Task C - Affect Analysis 20 marks

Use emotion and affect lexicon/s to perform sentiment analysis on each annotated category. The sentiment analysis should be performed on the entire corpus of an annotated category rather than analysing each tweet in the categories.

You are not required to limit your approach to any one theory of emotion or affect. You may pick any number and type/s of lexicon/s to support the analysis.

Task D - Text Classification 15 marks

Create a text classifier which can identify tweets with misinformation on Covid-19.

You are required to merge the annotation categories for tweets in a way that following class labels can be assigned during test and training phases: misinfo & no-misinfo. You may exclude some categories while merging the tweets into these two classes, but the decision should be supported with assumptions and empirical evidence.

You should test feature spaces with combinations of tokens, topics, named entities and concepts, and any other relevant engineered features.

Any feature spaces that you create to train the classifier should be informed by by previous tasks.

Reporting the tasks 5 marks

You are required to complement each of the previous tasks with a report style component/s. The reporting element should reflect and describe your analytical approach and related outcomes. The report style components in each task should tie up all the tasks into one well structured report.

Assignment Part 2 - Demonstration

Recorded demonstration video & Live Q&A 30 marks

The demonstration is used to test that you can evidence your approach to all tasks from Part 1.

So the demonstration and following Q&A will assess that you:

1. understand your code and
2. explain in detail the algorithms and methods utilised.

Useful hint: A well-integrated coding and reporting element from coursework can be very useful while you are demonstrating.

	80-100	70-79	60-69	50-59	40-49	30-39	0-29
	Exceptional First	**First**	**Upper 2nd**	**Lower 2nd**	**Third**	**Narrow Fail**	**Fail**
Analysis and Methods outline	Professional outline of analysis and methods used.	Detailed purpose of analysis and methods provided.	Adequate outline of analysis and methods provided.	Outline of analysis and methods provided but with some flaws.	Simple outline of analysis and methods provided, but lacking key detail.	Inadequate outline of analysis and methods provided.	No outline of analysis or methods provided.
Text mining & pre-processing	Sophisticated pre-processing of text data.	Comprehensive pre-processing of text data.	Adequate pre-processing of text data.	Pre-processing of text data is attempted but with some flaws.	Limited pre-processing of text data.	Inadequate pre-processing of text data.	No pre-processing of text data.
Text & Affect Analysis and Text classification	Unanticipated results and implementations presented. Appropriate, substantial, correct and sophisticated nature.	Comprehensive results and implementations, presented and employed well. Appropriate, substantial and correct.	Expected results and implementations presented. All appropriate, largely correct, with few flaws.	Not all expected results and implementations presented. All appropriate, largely correct, with few flaws	Few or simple results and implementations presented. Much appropriate material, but flawed.	Seriously flawed results or no implementation. Appropriate but seriously flawed material.	No results or implementation. Incorrect or inappropriate content.
Conclusions	Deep and critical understanding provided.	Thorough understanding shown.	Good understanding shown.	Key concepts generally understood.	Some evidence of understanding.	Little of superficial understanding shown.	No evidence of understanding.
Report	Like a publishable report, virtually error-free.	Like a publishable report with isolated minor errors.	Can be followed easily with very few errors.	Can be followed easily with some weaknesses.	Can be followed with difficulty.	Poor structure or containing significant errors.	Unstructured and with many errors.
Demonstration	Able to execute and explain the program clearly. Demonstrates an excellent level of understanding and explanation.	Able to execute and explain the program with no aid or guidance. Has a good level of understanding.	Able to explain the program with minor errors. Has a good but not complete understanding of the code.	Able to explain the program with some errors. Has some understanding of the code.	Able to explain the program with major errors. Makes an effort to explain the program but unable to do so.	Unable to explain in any detail the program that has been created Little awareness of the tasks or sections of code.	Unable to explain the program at all. No awareness of the purpose / location of any set tasks or sections of code.