ASSIGNMENT 2

## PLEASE READ ALL INSTRUCTIONS AND INFORMATION CAREFULLY.

This assignment contributes 50% to your final module mark. Please ensure that you retain a duplicate copy of your assignment work as a safeguard, in the unlikely event of your work being lost or corrupted online.

## THE FOLLOWING LEARNING OUTCOMES WILL BE ASSESSED:

* [LO2] Developed a critical understanding of Machine Learning, Data Mining and Data Analytics tools
* [LO3] Developed understanding of the professional, ethical, social and legal considerations involved in Data Mining and Data Analytics
* [LO4] Critically assess, choose and apply the appropriate Machine Learning, Data Mining and Data Analytics formalisms and tools to practical problems
* [LO5] Identify and assess data for the use of Data Mining and Data Analytics tools
* [LO6] Define, explain and interpret the results obtained from the practical application of Machine Learning, Data Mining and Data Analytics tools.

## IMPORTANT INFORMATION

You are required to submit your work within the bounds of the University Academic Regulations (see your Programme Guide). Plagiarism, paraphrasing and downloading large amounts of information from external sources, will not be tolerated and will be dealt with severely. The coursework submission for this module is largely based upon your own practice, but where you do use material from other sources, for example an occasional short quote, this should be duly referenced. It is important to note that your work **WILL BE SUBJECT TO CHECKS FOR ORIGINALITY**, which **WILL** include use of an electronic plagiarism detection service.

Where you are asked to submit an individual piece of work, the work must be entirely your own, except where you have correctly quoted, cited, and referenced external sources. The safety of your assessments is your responsibility. You must not permit another student access to your work at any time during the inception, design or development of your coursework submission and must take great care in this respect.

Where referencing is required, unless otherwise stated, the Harvard referencing system must be used (see your Programme Guide or university library website).

# **Aims**

This coursework focuses on the practical side of developing a Natural Language Processing Pipeline. The aim is to produce a pipeline for the given task, demonstrating good scientific technique and critical evaluation skills. As part of this you will be utilising industry standard tooling, and the work in this assignment requires the formulation of experiments to determine an appropriate solution.

You will be expected to write a report, and produce Python code (E.g Python Script or Python Notebook) supporting your findings.

# Task - Twitter Topic Classification

As part of the module you have looked at various techniques within Natural Language Processing, tying them together into a pipeline towards the supervised machine learning problem of classification; often referenced as a pipeline.

Your task is to produce a pipeline capable of solving a **multi-class classification task** in the form of a research project - whereby comparisons are made and assessed towards the task. For this you will follow the **CRISP-DM Methodology** (see Figure 2) covered in the module, evidencing your process in your report. Multiple approaches will be compared, and final evaluations and recommendations made.

For this assignment you are provided a JSON file containing 6443 entries which represent tweets from the social media platform Twitter, covering 6 topics. These tweets were gathered between 2019 and 2021 and were human-labelled using Amazon’s Mechanical Turk.

The categories of tweet cover a variety of topics, namely:

1. Arts & Culture
2. Business & Entrepreneurs
3. Pop Culture
4. Daily Life
5. Sports & Gaming
6. Science & Technology

You will be expected to read-in this dataset yourself using scikit-learn, pandas, and other library functionality; some of which we have covered in workshops. Table 1 outlines the Data Description for the provided dataset



Figure 1 - Example of a tweet on Social Media Platform, Twitter.

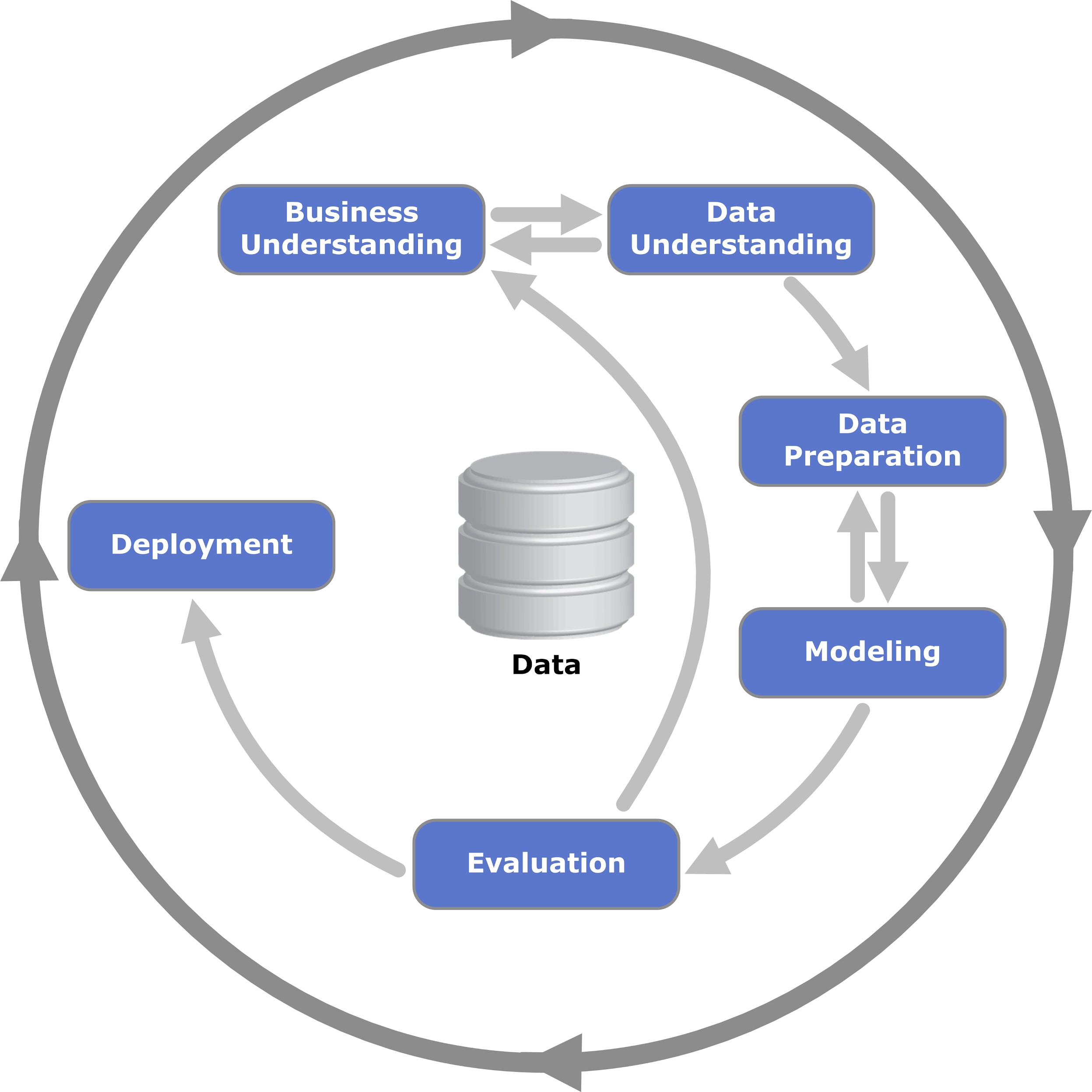


Figure 2 - The Standard CRISP-DM Data Mining Methodology.

Table 1 - Data Description for bespoke Twitter Dataset

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| text | Raw Content of the tweet. I.e The message posted by the user, including text, emojis, hashtags, and any @ mentions.  Contains pre-processed placeholders for certain content.  E.g {{URL}} represents a URL that has been replaced by that token; Specific user mentions are also tokenised {@HipHopDX@} for the user @HipHopDX |
| date | Date of the Tweet. Follows the YYYY-MM-DD format.  E.g 2021-08-01 |
| label | Numeric encoding of the label (As listed above), starting at 0.  E.g 2 |
| id | ID of the Tweet itself.  E.g 1421650405411405828 |
| label\_name | Human-readable name for the chosen category.  E.g “pop\_culture” |

**You are expected to appropriately read in the dataset, perform any required dataset partitioning, construct an appropriate pre-processing pipeline, train any relevant classification model(s) (e.g Naive Bayes, Logistic Regression, etc), and then evaluate it on the testing data.**

For this you should consider aspects of the NLP pipeline, including how we represent textual information and any techniques we’ve introduced which can assist us. An example of this would be the concept of vectorisation. This would be the representation of the text as numeric data, to then feed into appropriate classification models.

It is important to have a good methodological approach to this work, to follow the CRISP-DM methodology, and to ensure that experimental design is clear, concise, and well reported.

You may wish to consider starting with a simple baseline pipeline and model, and iterating upon it. It is expected you will use academic literature to inform your decision and choices and guide your search.

**Alongside the Python code for the proposed experiments, pipeline, and classification model, you will write a report ( 10 pages maximum ) outlining your solution**, the pipeline chosen, the classification model selection, any further processing of the dataset, as well as evaluative results. For the purposes of this report, you should carefully consider experimental design, showing comparisons between various different pipeline choices and classification models you’ve tried, using evaluative metrics to demonstrate an overall ‘good’ solution to the task.

# **Report Submission**

A report outlining the steps taken, aligned to CRISP-DM, towards solving the classification task. This should include any data reading and data pre-processing, the construction of your NLP pipeline and classification model, training, as well as any evaluations.

This report should include comparisons of various solution choices (E.g Vectorisation,

stop-words, regex filtering, classification model, etc). Therefore, it’s important to present the various experiments clearly, provide rationale behind the hypothesis (why you selected those elements to try), and an evaluative methodology to compare them (how will you select which solution is the best). If you found any ‘interesting’ results through your experiments, you should also present these.

The evaluation section of your report should evaluate the overall solution themselves towards the task, as well as form a conclusion on which selection (pipeline choices + classification model) is the best with appropriate justification. Commentary may be included in this section on the success of any pre-processing or data manipulation steps you’ve taken. For example, this could be the impact that filtering via regex had on the overall solution. This may be something to consider when setting out your various experiments.

**To assist with structuring you may find the following beneficial as heading suggestions:**

1. **Methodology** - Provide details on the methodology applied towards the task undertaken, providing rationale for these steps.

This should detail how you went from the raw data provided to the chosen model(s), choice of model, and how this methodology helps address the problem domain.

Evidence to support the following of this methodology should be presented, especially any cases which required moving backwards in the process to readdress issues.

1. **Results** - Presentation of any graphics, figures, tables, charts, etc which constitute as results. This could be the evaluative metrics presented from training of your solutions, it could also be interesting entries discovered through the process. There should be **no** discussion here of what results mean, only their presentation.

If using modified variants of the dataset (as part of your experiments / processing), these should be clearly identified in any tables with appropriate naming. The justification and description of such modification is not for this section.

Additional figures may be used as appropriate which can be in support of discussion points in the evaluation and discussion section, or as evidence for methodology following above.

1. **Evaluation & Discussion** - Outlining of an evaluative methodology; i.e How you went about evaluating your experiments and obtaining the results previously shown. Interpretation of your previously presented results (make frequent references back E.g “See Figure X” or

“Figure Y shows that…”. Discussion of these findings in relation to the overall task. How were these evaluated? Why was this selected? What metrics were used and why?

Discussion of the results should be presented with appropriate evidence and rationale. E.g Which is the best model, and why?

Consider each stage in the methodology and reflect on any improvements which could have been made. Could any techniques have been used which may have improved the performance of your results? Why?

1. **Future Works** - Based on the outcome of your experiments, and the conclusion you have made, what are potential avenues of investigation if you were to continue this, or have more time. Why are these worthwhile avenues to pursue? These suggestions should be grounded and supported by your own findings and relevant academic literature.
2. **References** - Academic references section.

# **Video Submission**

As part of this assignment you will present a video presentation demonstrating an understanding of the code you have written, and explain the NLP pipeline you have developed through the use of the CRISP-DM methodology.

The presentation is a maximum of 15 minutes, any work over that length will be ignored. Use the following as a template to structure your time.

Introduction (~2 minutes)

* Briefly introduce yourself and the purpose of the video, and provide a high-level overview of the task and the main objectives.

Code Walkthrough (~5 minutes)

* Walk through the main sections of your code, highlighting key functions and methods.
* Explain how the code implements the NLP pipeline, including data preprocessing, feature extraction, model training, and evaluation.
* Emphasise any unique or innovative aspects of your code; especially if it is technical in nature and beyond the scope of taught content.

Results and Evaluation (~5 minutes)

* Present the key results of your experiments, including performance metrics for various pipeline choices and classification models.
* Explain how you evaluated the results and determined the best solution.
* Discuss any interesting findings, challenges faced, or limitations of your approach.

Future Works and Conclusion (~3 minutes)

* Discuss potential improvements and future research directions based on your experiments and findings.
* Summarise the main takeaways from your work and reiterate the significance of your chosen solution.

# Deliverables & Submission Requirements

**Report** submission is required in PDF format with a **maximum of 10 pages**. You should ensure that the PDF is uploaded as-is, and not within a ZIP file or any other archive for the purposes of TurnItIn.

**Note**: The maximum page limit is 10, but submissions can be shorter. Any content beyond the 10-page limit will not be considered. The title page, table of contents, references, and appendices are not included in this limit.

Any submissions which do not adhere to guidelines in submitting in PDF will receive a 10-mark presentation penalty.

Work will be subject to any relevant over-length penalties

**Python Code** submission (.ipynb, .py, or a .zip of source code). There is no limit on the number of lines of code.

**Note:** Code must also be included as a single appendix item in the report (one unbroken entry), and in plain-text for the purposes of originality checks. Any work which does not adhere to this will receive a fixed deduction of 10 marks from the assignment.

**Video Presentation** submission (E.g .mp4 uploaded and linked to ReView). 15 minutes maximum. Video Presentation files should be of reasonable resolution (approx 1080p), of reasonable file-size (approx 500MB), and of a playable file format (E.g .mp4).

Links will **NOT** be accepted. The entire video presentation is to be uploaded to the video presentation submission zone. You may find it useful to record a short test-clip to verify you have appropriate settings for resolution and file-size prior to recording your presentation. Do not leave your upload of this component to the last minute, files should be fully uploaded into canvas before the deadline to be considered on-time. This is your responsibility.

# **Help with Referencing**

Whenever you need to refer the reader to the source of some information, e.g. a book/journal/academic paper/WWW address, provide a citation at that point within the main body of your report.

**Example 1**: ... as we are all now aware referencing is not trivial (Kendal, 2017)

Provide a reference list towards the end of your research paper (after your conclusions section but before any appendices) that contains:

* References, a list of books/journals/academic papers/URLs etc. that have been directly cited from within the report (see example citation above).
* Any material from which text, diagrams or specific ideas have been used, even if this has been presented in your own words, it must be cited within the main body of the paper and listed in the reference list. It is not enough to list this material in a bibliography.

**Example 2**: For Example 1, (using Harvard system) the reference list would contain the following:

Kendal S., 2017, Referencing standards, International Student Journal, Vol 55, Pages 25 – 30, Scotts Pub., ISBN 1-243567-89

This shows the authors, date published, title of paper (in single quotes), title of journal or conference (in italics), volume, page numbers, and publisher (ISBN desirable but not essential).

For further help see the following book which is available in the library:

* Cite Them Right: The Essential Guide to Referencing and Plagiarism by Richard Pears and Graham Shields

**END OF ASSESSMENT BRIEF**