Assessment Brief: Coursework 2023-24

Assessment Details

Course Title:	Theory and Applications of Data Analytics
Course Code:	LDSCI7236
Course Leader:	Ioannis Kypraios
Level:	7
First or Second Sitting:	First Sitting
Assessment Title:	Sentiment
Assessment Number:	AE2
Assessment Type:	Coding Assignment
Restrictions on Time/Length:	Code and 2500 word explanation
Individual/Group:	Individual
Assessment Weighting:	50%
Issue Date:	15 January 2024
Hand in Deadline:	12 April 2023
Planned Feedback Deadline:	28 calendar days after hand in deadline
File format accepted:	.ipynb, .zip
Mode of Submission:	Online
Anonymous Marking:	Yes

Assessment Task

Aims & Objectives

The assignment aims to give you practical experience in the design and implementation of a real-world machine learning application. You should be able to:

1. Train machine learning models with the provided data set and evaluate their statistical efficiency in order to find the best one for our particular application.

2. Fine-tune machine learning models, finding methodically the appropriate values for the hyper-parameters of a given model that yield the best statistical efficiency for our particular application
3. Deploy your model to a data set so that it can be used to predict the sentiment of reviews in real time.

Problem

The Stanford Sentiment Treebank[1] data set (SST-2) consists of ~68,000 sentences from movie reviews and annotations of their sentiment – 0 for negative reviews, 1 for positive reviews:

Sentence from review	*Sentiment*
"It's slow – very, very slow"	0
"It's a charming and often affecting journey"	1

SST-2 is one of GLUE benchmark tasks for natural language processing systems. The data set can be downloaded from gluebenchmark.com/tasks. Two files are of interest, train.tsv and dev.tsv, containing annotated reviews to be used for training and testing, respectively.

In this assignment, your task is to predict the sentiment of reviews in the test data set with high accuracy. Your report must include an analysis of:

1. **Exploratory data analysis.** Explore the distribution of positive and negative reviews in the SST-2 dataset, identify any data imbalances, and discuss potential implications for model training and evaluation. Apply appropriate visualisation of the data sets where necessary. **(30 marks)**
2. **The text vectorisation method of choice.** Compare and contrast the efficiency of at least two vector representations of movie reviews in the data set. Choose the most appropriate method based on its effectiveness in capturing the sentiment information. Consider providing explanations or justifications for your choices.**(20 marks)**
3. **The machine learning model of choice.** Compare and contrast the efficiency of at least two models for classification we have studied in class. Evaluate their performance in terms of accuracy, precision, recall, and F1-score. Discuss the strengths and weaknesses of each model and select the one that achieves the highest accuracy. **(15 marks)**
4. **The hyper-parameter values of choice.** Explore at least two hyper-parameters for a given model. Explain the rationale behind the selection of the explored hyperparameters by performing a systematic search and evaluation of different

[1] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of EMNLP, pp. 1631–1642. For more information, visit nlp.stanford.edu/sentiment.

hyperparameter values to optimise the model's performance. Report the best hyperparameter values that result in the highest accuracy. **(15 marks)**

5. **Model deployment.** Deploy the model that achieved the best accuracy to an evaluation data set so that it can be used to predict the sentiment of reviews in real time. Discuss the ethical implications of your solution, the scientific integrity of your design, and potential unintended consequences of your analysis more broadly. **(20 marks)**

Specification

1. Where requested a (sub)problem solution **should** be accompanied by:
   a. An explanation of the experiment(s) you find appropriate to answer the problem’s question(s).
   b. At least one visualisation – either a figure or a table.

   As a rule of thumb, your notebook’s text cells should explain ***what*** you want to achieve with your implementation; your inline code comments, on the other hand, should explain ***how*** you achieve it.
2. Every visualisation **must** be accompanied by a discussion of the data view illustrated.
3. You **must** provide appropriate references to any external resources you used or consulted. This excludes references to standard Python library functions. Use links where appropriate. Reference appropriately all the sources you have used in your implementation and discussion parts.
4. You **must** include a brief section (task 5 above) where you reflect on the ethical implications of your solution, the scientific integrity of your design, potential unintended consequences of your analysis, and the human and social context of machine learning more broadly. Your reflections should be ~1-2 paragraph.
5. You **must** explore (i) at least two vector representation methods, (ii) at least two machine learning models, and (iii) at least two hyper-parameters for a given model.

Submission

For this assignment:

- Prepare a single Jupyter Notebook file that includes all the code, text, printed outputs, and data visualisations.
- The notebook should be developed using Python 3 within a Jupyter environment, with well-commented code and appropriately structured text cells resembling a blog format.

- Avoid including personal name details in the notebook or its filename/directory.
- Remember to restart the kernel, thoroughly check the notebook for errors, and provide a working link to the dataset used in the analysis.
- Ensure that all code runs smoothly, and create a compelling narrative by integrating analysis, visualisations, and reflections.
- Your submission may include auxiliary Python files (*.py), where you packaged, e.g., some common functions and imported them in your notebook. In this case, submit a .zip file with all documents included.
- Do not include any dataset files in your submission.

Assessment Criteria

>70: There was evidence of the ability to perform all programming tasks correctly. The demonstration of the methods was excellent, coherent, well documented, and clearly explained.

60-69: There was evidence of ability to perform some programming tasks correctly. The demonstration of the methods is good, coherent, and reasonably detailed and explained.

50-59: There was evidence of ability to perform some programming tasks correctly, but the demonstration of the methods was limited, incoherent, not adequately documented and vaguely explained.

<49: Failure to solve the programming task in assignment. Methods were completely incorrect or absent.

Marking

The University uses two common assessment marking schemes – one for undergraduate and one for postgraduate – to mark all taught programmes leading to an award of the University.

More detailed information on the common assessment marking scheme and the criteria can be found in the Course Syllabus, available on the University’s VLE.

Learning Outcomes

On successful completion of the course, students should be able to:

Knowledge and Understanding

K1d Demonstrate a comprehensive understanding and knowledge of data science concepts and master their implementation in data analytic applications.

K2d Demonstrate critical awareness of feasible operations and transformation on data, and their relationships in current data processing pipelines.

K3d Demonstrate a degree of originality in plotting and visualising data in an effective manner.

K4d Critically review and identify key capabilities and limitations in data science practices, and propose directions for further innovation.

Subject-Specific Skills

S1d Critically evaluate basic data science concepts in their application for solving complex data problems.

S2d Critically evaluate the requirements and limitations of data transformations techniques to the chosen dataset.

S3d Demonstrate the ability to identify and implement efficient data science techniques to the area of application and produce clear and concise and well documented code.

S4d Identify appropriate data science practices within a professional, legal and ethical framework for addressing data management and use, security, equality, diversity and inclusion (EDI) and sustainability issues.

Transferable Skills

T1d Demonstrate initiative in leading and participating in teams for delivering data science projects in a timely manner and according to specification.

T2d Consistently display an excellent level of technical proficiency in written English and command of scholarly terminology, so as to be able to deal with complex issues in a sophisticated and systematic way.

T3d Demonstrate initiative in working independently, effectively, and to deadlines.

T4d Communicate effectively to both technical and non-technical audiences through oral presentations, software demonstrations, and written reports.

Accessing Feedback

Students can expect to receive feedback on all summative coursework within 28 calendar days of the submission deadline. The 28 calendar day deadline does not apply to work submitted late. Feedback can be accessed through the Turnitin assessment link on the course page. Further instructions on submitting an assessment and accessing feedback can be found on the University's VLE.

Late Submissions

Students are reminded to submit their assessment in the correct format and ahead of the published deadline. Deadlines are strict and Canvas uploads made remotely might not be immediate, we therefore strongly recommend that students upload their work to Canvas in good time before the deadline. If assessments are submitted late without approved Extenuating Circumstances, there are penalties:

- For assessments submitted up to two days late: any mark of 40% or higher will be capped at 40% for undergraduate students. Any mark of 50% or higher will be capped at 50% for postgraduate students. Any mark below 40% for undergraduate students and below 50% for postgraduate students, will stand.
- Students who do not submit their assessment within two days, and have no approved extenuating circumstances, are deemed not to have submitted and to have failed that assessment element. The mark recorded will be 0%.
- Late penalties are calculated differently for some types of portfolios. Please read the Assessment Brief of your portfolio carefully.

For further information, please refer to AQF7 Part C in the Academic Handbook.

Extenuating Circumstances

The University's Extenuating Circumstances (ECs) procedure is in place if there are genuine circumstances that may prevent a student submitting an assessment. If the EC application is successful, there will be no academic penalty for missing the published submission deadline.

Students are normally expected to apply for ECs in advance of the assessment deadline. Students may apply for consideration of ECs retrospectively if they can provide evidence that they could not have done so in advance of the deadline. All applications for ECs must be supported by independent evidence.

Students are reminded that the ECs procedure covers only short-term issues (within 21 days leading to the submission deadline) and that if they experience longer-term matters that impact on learning then they must contact Student Support and Development for advice.

Under the Extenuating Circumstances Policy, students may defer an assessed element on only one occasion and may request an extension on a maximum of two occasions.

For further information, please refer to the Extenuating Circumstances Policy in the Academic Handbook.

Academic Misconduct

Any submission must be a student's own work and, where facts or ideas have been used from other sources, these sources must be appropriately referenced. The Academic Misconduct Policy includes the definitions of all practices that will be deemed to constitute academic misconduct. This includes the use of artificial intelligence (AI) where not expressly permitted within the assessment brief, or in a manner other than specified. Students should check this policy before submitting their work. Students suspected of committing Academic Misconduct will face action under the Policy. Where students are found to have committed an offence they will be subject to sanction, which may include failing an assessment, failing a course or being dismissed from the University depending upon the severity of the offence committed. For further information, please refer to the Academic Misconduct Policy in the Academic Handbook.

Version History

<table>
<tr><th colspan="6">Title: Assessment Brief Template
Approved by: The Quality Team</th></tr>
<tr><th>Version number</th><th>Date approved</th><th>Date published</th><th>Owner</th><th>Location</th><th>Proposed next review date</th></tr>
<tr><td>4.0</td><td>March 2023</td><td>March 2023</td><td>Registrar</td><td>VLE/ Faculty Resources Page</td><td>March 2024</td></tr>
<tr><td>3.0</td><td>August 2022</td><td>August 2022</td><td>Registrar</td><td>VLE, Faculty Resources Page</td><td>July 2023</td></tr>
<tr><td>2.3</td><td>December 2021</td><td>December 2021</td><td>Registrar</td><td>VLE</td><td>August 2022</td></tr>
<tr><td>2.2</td><td>August 2021</td><td>August 2021</td><td>Registrar</td><td>VLE</td><td>August 2022</td></tr>
<tr><td>2.1</td><td>September 2020</td><td>September 2020</td><td>Registrar</td><td>VLE</td><td>August 2021</td></tr>
<tr><td>2.0</td><td>September 2020</td><td>September 2020</td><td>Registrar</td><td>VLE</td><td>August 2021</td></tr>
<tr><td>1.0</td><td>August 2019</td><td>August 2019</td><td>Registrar</td><td>VLE</td><td>August 2020</td></tr>
<tr><td colspan="6"></td></tr>
<tr><td>Referenced documents</td><td colspan="5">AQF7 Academic Regulations for Taught Awards; Extenuating Circumstances Policy; Academic Misconduct Policy; Course Syllabus</td></tr>
<tr><td>External Reference Point(s)</td><td colspan="5">UK Quality Code Theme: Assessment</td></tr>
</table>