IT2394 TEXT AND SOCIAL ANALYTICS PROJECT

Project Guide

AY2023 Semester 2

Table of Contents

1 Introduction

1.1 WIU synopsis

With a large amount of data generated by social media platforms, text analytics provides an effective way to conduct market research on product branding and reputation.

Through this WIU, learners will demonstrate their competencies in developing a text & social analytics project in teams by analysing textual data such as customer feedback, social media conversations and textual documents to understand customer experiences, preferences and needs with consideration of diverse cultural differences.

Learners will be competent to perform data processing operations in NoSQL database and common text analytics tasks such as entity extraction, text categorisation, clustering, association and trend analysis in the textual data that comply with data and privacy ethics.

Learners will also be competent in undertaking a technical lead role to drive team discussions in developing data stories for an effective narrative and visual representation of their text & social analytics project.

1.2 WIU learning outcomes

At the end of this WIU, learners will be able to:

1. Collect and analyse data on customer experience and behaviour using research frameworks and historical data that comply with data and privacy ethics.
2. Perform data processing and aggregation operations on textual data in the NoSQL database using data manipulation language to address the business needs.
3. Extract key topics, its association and subjective information in textual conversations with consideration of diverse cultural differences.
4. Work collaboratively in a team to develop dashboards using data storytelling approach for an effective narrative and visual representation of customer profiles and social sentiments.

2 Project Scope & Scenario

2.1 Project Background & Scope

About 80% of enterprise-relevant data is in unstructured or semi-structured format. These include emails, documents, surveys, feedback forms, warranty claims, contact-centre notes and transcripts, web pages, news, data from social media, audios, videos and many more. A predominant amount of such data is available as text.

In order to gain competitive edge in the market, businesses and organisations are finding a growing need to expand their analysis scope to cover text data, especially with regards to customer feedback and social media data. This is so that critical insights can be uncovered to support business decision making and process improvement.

You are a group of creative, innovative, and energetic programmers who intend to grab the text from the internet with aim of analysing textual data such as customer feedback, social media conversations and textual documents to understand customer experiences, preferences and needs with consideration of diverse cultural differences.

Each team shall come up with the project proposal, collect textual data using web scraping with python, clean data with text pre-processing and build text classification model.

2.2 Project Requirements

Key Task 1: Project Proposal

1. As a team, discuss and identify a business problem and objectives that is related to text classification area.

   For example, choose a particular product or service (i.e. book or restaurant). Scrape the reviews for that product or service and construct it as a dataset. Create a text classification model to detect whether a review expresses a positive or negative opinion.

2. Explore textual data from Internet to collect your own dataset.

3. Describe the purpose of the project. What is the target variable you are trying to predict?

4. Plan on how to carry out the project. You should identify the main and subtasks and delegate the tasks to each of the members.

5. Specify the techniques and the python library to be used.

Key Task 2: Collect text data using web scraping

1. Once the team is formed and business objectives are defined, you will need to collect data that are relevant to achieve the objects. Scrape text data on the internet programmatically using python. The techniques to collect data should comply with data protection and privacy ethics.

2. Identify proper information for the target variable. If you need to annotate the data manually, be mindful of the time spent.

3. Each team member should choose a different website to scrape data.

4. The team collects the data from team members and combines them into one CSV format file.

5. The team splits the CSV file into two CSV files for submission: one is the raw dataset for training and evaluation purposes, and the other is the test dataset for accurate assessment.

Key Task 3: Perform data cleansing

1. In this task, each team member is required to apply text pre-processing technique from IT2391 Natural Language Processing to clean the unstructured textual data in the raw dataset, preparing it for modelling.

2. Store the cleaned data in a separate csv file.

Key Task 4: Text classification modelling

1. In the modelling task, each team member should apply relevant text classification modelling techniques on the data. It is important to select the relevant modelling techniques.

2. Each team member should build and train multiple models.

3. Each team member should assess and compare the trained models to identify the best possible outcome. Apply the best model on test dataset.

4. The team consolidates the outcome of the test dataset from team members and reflects on the challenges you have encountered and learnings you have gained while working on this project.

Key Task 5: Individual project demonstration and interview

1. Finally, evaluate your final model and assess if it meets the business objectives. Demonstrate the project. Provide useful interpretation of the outcome and recommendation.

2. You would also have to complete a mock job interview individually and highlight your suitability for the job by showcasing your skills, knowledge and aptitude.

2.3 Roles & Responsibilities

Each project group should consist of **4 students**. Each group must elect a group leader.

Module Tutor

- Provide feedback to project teams' proposals and approach.

- Advise the project group during project development, including the task allocation, development plan, tools used, etc.

- Provide inputs from the user perspective to the project team.

- Act as technical advisors (module tutors are **NOT** expected to debug programs or solve problems for students).

Group Leader

- Maintain the project schedule and documentation, and record attendance at meetings.

- Coordinate and assign work among members.

- Arrange meetings with the module tutor. Leader is required to book the meeting rooms if meeting is held outside of the lab/tutorial hours.

- Communicate with the tutor on behalf of the team to discuss any project-related matters.

Group Members

- Carry out allocated tasks.

- Be punctual and attend all lessons and project meetings.

- Cooperate with the rest of the group during project development.

2.4 Project Schedule & Submission

This project will be completed over 8 weeks and will adopt the Cross Industry Standard Process for Data Mining Initiative (CRISP-DM) methodology. Please refer to the following six key phases from CRISP-DM and the corresponding schedule. The detailed lesson plan can be found in Brightspace.

CRISP-DM	Weekly Schedule & Submissions
1. Business Understanding a. Identify and describe business problem b. Determine project goals	**Week 9**: Team forming & project goal identification **Week 12**: Identify data source on internet **Week 13**: Prepare Project Proposal
2. Data Understanding a. Describe data b. Explore data c. Verify data quality	***Project Proposal Submission (15%, group)*** *(refer to Annex E)* *[due on 15th Jan 2023, Sunday, 2359]*
3. Data Preparation a. Scrape data b. Construct/integrate/format data c. Clean data	**Week13-15**: Technical Review - Data Preparation • ***Data Scraping Jupyter Notebook submission (15%, individual)*** • ***Raw dataset and test dataset (5%, group)*** • ***Data Preparation Jupyter Notebook submission (15%, individual)*** *[due on 29th Jan 2023, Sunday, 2359]*
4. Modelling a. Select modelling technique b. Build model c. Assess model 5. Evaluation a. Evaluate models b. Review process	**Week 16-17**: Technical Review - Models Creation, Modelling Evaluation • ***Text classification Jupyter notebook submission (15%, individual)*** • ***Model evaluation report (5%, group)*** *[due on 12th Feb 2023, Sunday, 2359]*
6. Project Demonstration	**Week 18**: Project Demonstration and Mock Job Interview • ***Project demonstration (10%, individual)*** • ***Mock Job Interview (20%, individual)***

All submissions are to be submitted via Brightspace. Late submissions will be penalised.

3 WIU Assessment Components

Assessment Component	Group	Individual	Total
Project Proposal	15%	-	15%
Technical Review - Data Preparation	5%	30%	35%
Technical Review - Modelling and Evaluation	5%	15%	20%
Project Demonstration		10%	10%
Mock Job Interview		20%	20%
Total			100%

The assessment will be done through:
- Continuous assessment by module tutors throughout the duration of the project.
- Assessment by independent markers during project demonstration and mock job interview.

Annex A: Project Assessment Rubrics

Project Proposal (15%)

Criteria	Proficient	Competent	Functional	Developing	Not Competent
Innovation and Pitching (5%, group)	• Radical innovation, indicating creativity and originality with a significant contribution to business needs. • Gained support for the idea from the resource owner using persuasion technique	• Architectural innovation, indicating creativity with reasonable contributing to business needs • Explained the value of an idea to resource owners using a persuasion technique to get support	• Modular type of innovation, indicating creativity, but with little contribution to business needs. • Conveyed the intent of an idea to resource owners using persuasion technique to get support	• Incremental innovation, indicating low creativity and originalityAttempted to convey an idea to resource owners using persuasion technique to get support	• No innovation, indicating very low creativity and originality • Insufficient evidence that the pitch had clear purpose
Proposal, Co-creating and Project Planning (10%, group)	• Proposed business needs and solution exhibits high level of empathy for user, and incorporate feasible, realistic, and practical workflow that value add to business needs and processes. • Concisely and clearly describe the business scenarios and project goal of using text analytics to solve the problem • Prepared an action plan to achieve key milestones for the project • Engaged appropriate stakeholders in project development with aligned common goals	• Proposed business needs and solution exhibits good empathy for user and incorporated mostly feasible, realistic workflows that requires some fine tuning before it can value add to business needs and processes • Describe the business scenarios and project goal of using analytics to solve the problems • Prepared a roadmap to execute the project based on key milestones • Engaged appropriate stakeholders in project development with aligned common goals	• Proposed functionality exhibits some levels of empathy for user and incorporate workflow that requires some enhancement before it can value add to business needs and processes. • Describe the business scenarios and project goal of using analytics to solve the problems • Determined milestones with consideration of desired outcomes for each stage of the project development • Engaged appropriate stakeholders in project development ideation	• Proposed functionality exhibits low level of empathy for user and incorporate workflow that requires substantial enhancement before it can value add to business needs and processes. • Confusing description on the business scenarios and project goal of using analytics to solve the problems • Identified desired outcomes for each stage of the project development • Engaged appropriate stakeholders in project development discussions	• Proposed functionality exhibits little empathy for user and incorporate workflow that requires substantial enhancement before it can value add to business needs and processes. • No description on the business scenarios and project goal of using analytics to solve the problems • Insufficient evidence that a project plan with key milestones was formulated • Insufficient evidence that the team worked with stakeholders to develop the project

NANYANG POLYTECHNIC | School of Information Technology

Technical Review - Data Preparation (35%)

Criteria	Proficient	Competent	Functional	Developing	Not Competent
Data Scraping (15%, individual)	• All the code cells in the submitted notebook can be executed without any error • Able to obtain raw dataset with sufficient data in csv format from submitted notebook in a reasonable response time • Detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with excellent value-added insight to the	• At least 90% of code cells in the submitted notebook can be executed without any error • Able to obtain raw dataset with sufficient data in csv format from submitted notebook in a reasonable response time • Detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with good value-added insight to the outcome of code	• At least 70% of code cells in the submitted notebook can be executed without any error • Able to obtain raw dataset with sufficient data in csv format from submitted notebook in a reasonable response time • Somewhat detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with substantial value-added insight to the outcome of code	• At least 50% of code cells in the submitted notebook can be executed without any error • Able to obtain raw dataset but with insufficient data in csv format from submitted notebook • Adequate detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with adequate value-added insight to the outcome of code	• Less than 50% of code cells in the submitted notebook can be executed without any error • Able to obtain raw dataset but with insufficient data in csv format from submitted notebook • Insufficient detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with inadequate value-added insight to the outcome of code
Data Preparation, Data Management, and Iteration (15%, individual)	• Able to identify and apply comprehensive pre-processing techniques and feature engineering to text data • All the code cells in the submitted notebook can be executed without any error • Detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with excellent value-added insight to the outcome of code • Interpreted the insights obtained through systematic data analysis • Improved the usability based on prototyping and testing	• Able to identify and apply all necessary pre-processing techniques and feature engineering to text data • At least 90% of code cells in the submitted notebook can be executed without any error • Detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with good value-added insight to the outcome of code • Identified relational/co-relational insights through systematic data analysis. • Improved specific features on the prototype based on prototyping and testing	• Able to identify and apply most of the necessary pre-processing techniques and feature engineering to text data • At least 70% of code cells in the submitted notebook can be executed without any error • Somewhat detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with substantial value-added insight to the outcome of code • Identified patterns and trends through systematic data analysis • Made improvements to the ideas and conceptual model with considerations that included user feedback	• Able to identify and apply some of the necessary pre-processing techniques and feature engineering to text data • At least 50% of code cells in the submitted notebook can be executed without any error • Adequate detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with adequate value-added insight to the outcome of code • Attempted to generate insights through data analysis • Made improvements to the ideas with considerations that included user feedback	• Not able to apply any of the pre-processing techniques to text data • Less than 50% of code cells in the submitted notebook can be executed without any error • Insufficient detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with inadequate value-added insight to the outcome of code • Insufficient evidence that data was analyzed for the project • Insufficient evidence that ideas were improved using user feedback
Raw Dataset (5%, group)	Raw dataset submitted with large and high relevant information for text data analysis	Raw dataset submitted with sufficient and high relevant information for text data analysis	Raw dataset submitted with acceptable and high relevant information for text data analysis	Raw dataset submitted with acceptable and some relevant information for text data analysis	Raw dataset submitted with insufficient relevant information for text data analysis

Technical Review - Modelling, Evaluation (20%)

Criteria	Proficient	Competent	Functional	Developing	Not Competent
Prototyping and testing (15%, Individual)	• Models are built with comprehensive design and consideration • Models were evaluated with detailed criteria and assessment • All the code cells in the submitted notebook can be executed without any error • Detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with excellent value-added insight to the outcome of code • Tested the usefulness of the idea on users using prototypes	• Models are built with all appropriate design and descent amount of consideration • Models were evaluated with all appropriate criteria and assessment • At least 90% of code cells in the submitted notebook can be executed without any error • Detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with good value-added insight to the outcome of code • Investigated the features of the idea to users using functional prototypes	• Models are built with some appropriate design and some consideration • Models were evaluated with some appropriate criteria and assessment • At least 70% of code cells in the submitted notebook can be executed without any error • Somewhat detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with substantial value-added insight to the outcome of code • Demonstrated the features of the idea to users using functional prototype	• Models are built with appropriate design but not given enough consideration • Models were evaluated with simple criteria and assessment • At least 50% of code cells in the submitted notebook can be executed without any error • Adequate detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with adequate value-added insight to the outcome of code • Demonstrate the concepts of the idea to users using concept prototype	• Models are built with inappropriate design and/or some consideration • Models were evaluated with inappropriate criteria and/or assessment • Less than 50% of code cells in the submitted notebook can be executed without any error • Insufficient detailed examine, analysis, describe of solution/purpose in markdown cells of submitted notebook with inadequate value-added insight to the outcome of code • Insufficient evidence that a prototype was created to test ideas
Model Evaluation Report (5%, group)	• Consolidate the outcome of test dataset from all the team members. • Provide the relevant and appropriation explanation on the model outcome comparison. • Shows strong evidence of reasoned reflection and depth. • Adequate justifications given and application of learning clearly described.	• Consolidate the outcome of test dataset from all the team members. • Provide the relevant and appropriation explanation on the model outcome comparison. • Shows evidence of reasoned reflection but lack depth. • Some justifications given and application of learning clearly described.	• Consolidate the outcome of test dataset from most of the team members. • Provide some relevant explanation on the model outcome comparison. • Shows little evidence of reasoned reflection. • Few justifications given and application of learning clearly described	• Consolidate the outcome of test dataset from some of the team members. • Lacks reflection and depth. • No justifications given and application of learning clearly described	• No report submitted

Project Demonstration (10%)

Criteria	Proficient	Competent	Functional	Developing	Not Competent
Project demonstration, Communication and Leading (10%, Individual)	• Excellent solution demonstration with no error • Able to relate model results to real world problem and make novel recommendation • Conveyed and discussed ideas with stakeholders using communication methods and tools effectively and garnered support • Led and managed the team to develop the project	• Good solution demonstration with no error • Able to relate model results to real world problem and make useful recommendation • Conveyed and discussed ideas with stakeholders using communication methods and tools • Contributed to managing some parts of the project	• Adequate solution demonstration with 2 or less errors. • Able to relate model results to real world problem and make useful recommendations • Discussed ideas within the project team using communication methods and tools • Guided team members to achieve some parts of the project	• Solution demonstration with errors that required the solution to restart for the continuation of solution demonstration. • Able to relate model results to real world problem and make simple recommendations • attempted to discuss ideas within the project team using communication methods and tools • Supported the leader to help the team perform better	• Solution demonstration with a lot of errors that required the solution to restart for the continuation of solution demonstration. • Unable to relate model results to real world problem to make any recommendation • Insufficient evidence that development of ideas was discussed within the project team • Insufficient evidence or appreciation and support for leadership

Mock Job Interview (20%)

Criteria	Proficient	Competent	Functional	Developing	Not Competent
Quality of Answers (8%, Individual)	• Strong, convincing answers to demonstrate: - Ability to do the job - Ability to fit in - Stress tolerance	• Convincing answers to support: - Ability to do the job - Ability to fit in - Stress tolerance	• Some convincing answers to support: - Ability to do the job - Ability to fit in - Stress tolerance	• Adequate answers to support: - Ability to do the job - Ability to fit in - Stress tolerance	• Answers are weak and not forthcoming • Interviewers need to probe for answers
Quality of Questions (4%, Individual)	• Questions show keen interest in the job and the willingness to contribute to the company • Questions draw attention to strengths, suitability and the willingness to fit in	• Questions show keen interest in the job • Questions draw some attention to strengths, suitability and the willingness to fit in	• Questions show interest in the job	• Questions show some interest in the job	• No questions asked, or questions asked are inappropriate
Confidence and Enthusiasm (6%, Individual)	• Responds to remarks and/or questions with enthusiasm throughout • Demonstrates much confidence and highlights own suitability with much passion and pride • Uses paraverbal and body language to engage the interviewer all the time	• Responds to remarks and questions with enthusiasm most of the time • Demonstrates good confidence and highlights own suitability with some passion and pride • Uses paraverbal and/or body language to engage the interviewer most of the time	• Responds to remarks and questions with some enthusiasm. Some responses are verbalisation of memorised script • Demonstrates confidence and highlights own suitability at times • Uses paraverbal and/or body language to engage the interviewer at times	• Responds to remarks and questions with signs of enthusiasm. Many responses are verbalisation of memorised script • Demonstrates confidence and highlights own suitability at times but may show unease at times • Uses paraverbal and/or body language and attempts to engage the interviewer at times	• Gives little or no response even when prompts are given, or only verbalise memorised script • Shows a critical lack of confidence and much unease throughout • Shows a critical lack of interest (e.g. no eye-contact throughout, not paying attention, etc.) in being interviewed
Attire (2%, Individual)	• Appropriately attired • Well groomed		• Attire may be inappropriate or too casual • Improvements needed in terms of grooming		

NANYANG POLYTECHNIC | School of Information Technology

Annex B – Form for Submitting Team Organisation
(To be completed by the Team)

Module Group: _____

Team Name (optional): _____

S/n	Name	Admin No.
1.*		
2.		
3.		

*denotes the Team Leader

Official (Open)

Annex C – Project Proposal (15%)
(To be completed by the Team)

WIU **Group:**	IT2394-0*
Team Name:	
Team Members: **(*denotes the Team Leader)**	1. <Student Name> (<Admin No>)* 2. 3.

The goal of the project proposal is for module tutor to give you early feedback on your project plans and to determine whether the scope of your project is appropriate. The project proposal should include the following information:

- A description of the business area (e.g. Product, Service, etc.) you wish to investigate

- A description of background, business scenario and business objectives

- List Down the main and sub tasks for the team. Delegation of tasks/responsibility.

- A description of how you will collect your own dataset. Provide a hyperlink of web where you plan to scrape the text.

Annex C – Project Proposal (15%)

Annex D – Mock Job Interview (20%)
(To be completed individually)

<u>Project Brief</u>

This ICA assesses your ability to engage interviewers and leave a positive impression at a job interview.

This assignment is in the context of you being interviewed for an internship position as a Data Analyst in Lazada.

The interviewers are role-played by your tutors. You are expected to engage the interviewers effectively by answering questions well to highlight your suitability for the job, as well as asking questions that show good interest and willingness to contribute to the company.

Tips for Preparation

Conduct adequate research on the company and prepare answers to questions which the interviewers are likely to ask. Your answers should help you highlight your knowledge, skills and ability to do the job, your value-add to your work team, and your ability to adapt to the company's culture.

Prepare questions based on the company profile and the requirements of the job position. Your questions should show that you are enthusiastic in joining the company and that you are well-prepared for the job position.

You are also required to come appropriately attired for the interview to project a professional and positive image of yourself.