Group Assignment: Natural Language Processing for Customer Insight Analysis

CSC6202 - Assignment 3

Due: 26 April 2024

Objective

As data scientists, a key competency is the ability to lead an AI enterprise workflow. Therefore, data scientists must possess relevant technical knowledge, along with domain expertise and skills in teamwork and communication.

For this project, you will leverage NLP techniques to analyse Amazon Customer Reviews for thematic insights, perform a sentiment assessment, and develop a basic Question-Answering (QA) system. You will draw on concepts and techniques from lectures, workshops and your own independent study. The aim of this project is to provide an explainable system to stakeholders that is designed ethically for potential deployment. It will provide actionable product improvement recommendations to improve customer interactions for a real world scenario.

In a team of THREE (3) students, you must produce and deliver:

1. A **Product** in the form of a QA system using an NLP development process (24%)
2. **Slides** that function as a report to stakeholders, explaining the project and the product (12%) (this contains an ethical assessment - see below)
3. A team **Presentation** using the slides, delivered live in the last workshop of the trimester or through a prerecorded video (10%)

The team must also conduct an **ethical assessment** of the project and product. The ethical assessment will contain issues documented during the project lifecycle, discuss any ethics the company must consider for deployment, and provide a review plan (See Task 4 in this specification). The ethical assessment is a **section** in your presentation slides and is worth 4%.

This assignment will achieve Course Learning Outcomes 1-6 listed in the CSC6202 handbook.

Details

Due: 26th April 2024
Weightage: 50%

Email Group Sheet to Course Coordinator

After forming your team of 3, complete the Assignment 3 group sheet available in the Assignment 3 section on StudyDesk. This includes a brief social contract. The social contract should be completed by all three members and is a list of behavioural principles the group agrees to abide by during the project. One team member should email the completed group sheet to the course coordinator prior to the project so your StudyDesk group submission system can be activated.

Problem

A mid-sized e-commerce company specialising in electronics seeks to understand customer perspectives on its product range. Their issue is they receive many enquiries each day that they cannot consistently answer, and they feel they are losing business to competitors. They have heard about the use of NLP in automated QA systems, but are concerned with recent news reports around privacy and ethical issues associated with AI.

Your team is tasked with developing a solution to present to this company.

Dataset

- The Amazon Customer Reviews Dataset is located in the Assignment 3 specifications folder. This contains qualitative customer feedback.
- From this dataset, your group's task is to extract meaningful insights through thematic analysis, determine general sentiment, and build a QA system.

1 Product

Task 1: Thematic Analysis for Product Improvement

Analyse the reviews to identify common themes.

Process

- Identify recurring themes in the reviews and categorise them using the below two approaches.
    1. Analyse the most common or frequent words in the reviews to gain quick insights into frequent topics or concerns highlighted by the customers.
    2. Utilise topic modelling methodologies, such as Latent Dirichlet Allocation (LDA), which systematically identifies and classifies underlying topics within the textual content.
- Evaluate both methodologies to determine the most suitable approach for this project, and justify the selection.

In your slides:

Cover your methodology and how it has been applied to the dataset and development of the product. Exhibit the results of your evaluation in your presentation.

Task 2: Sentiment Analysis

Develop an NLP model to categorise reviews into sentiments.

Process

- Preprocess the data, removing any irrelevant information or noise.
- Train a sentiment analysis model using an appropriate NLP library.
- Validate the model with a separate test set and document its performance.
- Conduct a statistical analysis to present the distribution of sentiments (positive, negative, neutral) across various products and previously identified themes.
- Examine the relationship between the customer ratings and the sentiments extracted through sentiment analysis to distinguish any patterns or discrepancies.

In your slides:

Provide a brief performance evaluation of the sentiment analysis model. Report your observations of the relationship between the customer ratings and the sentiments from the dataset. Discuss any potential biases and ethical considerations.

Task 3: Simple Question Answering System

Create a QA system to answer frequent customer inquiries based on review content automatically.

Process

- Collaborate with your team to compile a list of common questions customers might ask about the products. For example, ”What do customers think about the durability of the product?” or ”Are there any complaints about the product’s scent?”
- For each question, identify a set of keywords or phrases that are likely to be relevant to the answer. For example, for a question about durability, keywords might include ”durability,” ”long-lasting,” ”broke,” ”quality,” etc.
- Write a function to search the reviews for sentences containing the keywords associated with each question. This function can use basic string matching or regular expressions to find matches.
- If a sentence contains keywords related to a question, consider that sentence part of the answer to the question.
- Aggregate all sentences found for each question as the ”answer” from the system. If multiple sentences from different reviews address the question, include them all, giving users a broader view of customer opinions.
- Evaluate the system’s performance by checking if the retrieved sentences provide relevant information answering the questions. Discuss the system’s limitations and potential areas for improvement.

Task 4: Ethical Assessment for the Company

Ethics should be considered through the project lifecycle, including development of the product and use of data, and as an overall assessment of the product before deployment. This should be a section in your presentation (i.e. on the slides and delivered during the talk). You should:

- Discuss how any personal information was anonymised and how data was handled in compliance with relevant privacy laws.
- Provide key ethical insights the company must consider before the system is deployed, such as any potential biases and ethical considerations.
- Include a plan for regular reviews of the model's performance and impact over time.

Deliverables

1. Product (24%)

- Submit a well-documented Jupyter Notebook that details all the tasks implemented in this assignment. Comments should indicate where each group member contributed to a component (i.e. at the end of a comment, put either the student name or initials - multiple students can be attributed to parts).
- The provided dataset from StudyDesk does not need to be included, but the Notebook must run independently for the marker. Do NOT change the name of the dataset in your implementation.

2. Slides (12%)

- Submit presentation slides for a presentation between 12-15 minutes. Slides can be designed using programs such as Powerpoint or LaTeX (i.e. be submitted as .ppt or .pdf).
- The slides serve as a visual report to the company. Adjust your concepts and language to the nature of the project (e.g. make reference to the dataset and objectives) so the audience understands how you applied the NLP pipeline to the problem.
- Where applicable, use visualisations (e.g. plots, graphs, tables, etc) to convey messages effectively and use relevant samples from the project to highlight key ideas. All visualisations, numbers, and samples should be produced by your Notebook (you can copy metrics into tables for easier presentation).
- The structure of the slides should reflect the project. Some sections to consider are:
  - Problem (your understanding of the issue(s) the company faces)
  - Methodology (application of the NLP pipeline)
  - Results (such as analysis of the e-commerce product reviews, discussion of the themes identified, the sentiment of the customers towards the products, etc)
  - Question-Answering system

  - Ethical assessment (addressing Task 4 in this specification)
  - etc

You may include Appendices slides with supplementary information (e.g. extended metrics, additional visualisations, etc) that could not fit within the 12-15 minutes presentation, but is important for the company to know. These Appendices slides should simply be referenced in the main slides and during the presentation.

3. Presentation of Slides (10%)

- A presentation using the submitted slides that is at least 12 minutes (no longer than 15). It should have an equitable contribution from all team members.
- There will be two options. A team can choose to:
  - Deliver the presentation live as a group in the final workshop of the trimester on campus. This will be recorded by the facilitator for you to submit on the due date; OR
  - Submit a prerecorded video involving all group members to StudyDesk.
- If the video is too large to upload, the video can be made available via a hosting link (e.g. Panopto, SharePoint, Google Drive, Youtube, etc). Put the link to the video on the title slide or submit it in a txt file. It MUST be accessible to the marker (e.g. do not put it behind a password - if hosted on Youtube, make it unlisted).

4. Ethical Assessment (4%)

- This is a section in your Slides that reports on the requirements of Task 4 in this specification.
- Do not make the ethical assessment an Appendix item in the slides. This section must be in your main slides and presented to obtain marks for it.

Submission Guidelines

- Submit your code, slides and presentation recording through the course's StudyDesk. Submit each as a separate file (include the link to the video on the title slide or as a txt file). All items can be submitted by the due date.
- Ensure the names, student IDs, and assignment details are indicated in the code and report for each student.
- Late submissions incur a penalty of 5% of the Mark awarded to the student(s), per Calendar Day late.
- References are not compulsory, but are beneficial if you consulted any sources for your project.

Code Quality Best Practices

To obtain maximum marks, the code should adhere to best practices:

- **Readability:** Code should be easy to read and understand, even for someone who didn't write it. This includes using meaningful variable and function names that convey purpose without needing additional comments, and clear segmentation and control structures.
- **Consistency:** Consistent coding style makes the code more predictable and easier to read. This includes consistent naming conventions, indentation, and spacing.
- **Comments and Documentation:** Comments should be used to explain the ”why” behind complex logic, not the ”what”. Documentation should provide a clear overview of the code's functionality, including how to run it and what each part does. In this assessment, indicate sections where team members made a contribution (using either names or initials). Multiple team members could work on the same part if the process is collaborative.

Presentation Best Practices

- **Clarity and Conciseness:** The slides should convey information clearly and concisely, without unnecessary jargon or overly complex sentences. Every section and bullet point should have a clear purpose.
- **Structure and Organisation:** The slides should have a logical structure, including an introduction, methodology, results, discussion, and conclusion. Use headings to guide the reader through the content.
- **Visual Aids:** Incorporate charts, graphs, and tables where appropriate to visually represent data and findings. Make sure each visual aid is clearly labelled.
- **Critical Analysis:** Beyond presenting findings, the presentation should analyse and interpret results, discuss their implications, and how they relate to the project objectives.
- **Proofreading and Editing:** Before final submission, proofread to correct any spelling, grammar, or punctuation errors, ensuring professionalism and readability.

Academic Misconduct

Future AI and data scientists are expected to adhere to a code of conduct and approaches to ethics. Remember, we design systems to improve large scale problems, not substitute for the competencies required in our profession. UniSQ has zero tolerance for academic misconduct, including plagiarism and collusion.

- Plagiarism is presenting someone else's work as if you wrote it yourself.
- Collusion is a specific type of cheating that occurs when two or more groups exceed a permitted level of collaboration on a piece of assessment. Students are expected to work within their teams and not share their work with other groups.

- Identical layout, identical mistakes, identical argument and identical presentation in students' assignments are evidence of collusion.

- The University of Southern Queensland has strict policies on the use of Artificial Intelligence in the production of reports, code and presentations. If the use of AI is detected by the system, this will escalate to the Academic Integrity Unit and delay your results.

- Refer to UniSQ Policy Academic Misconduct for further details.

Marking Criteria

Product (24 marks)

- Excellent
  - Thematic Analysis: Utilises advanced NLP techniques (e.g., LDA) and most common or frequent words in the reviews with a thorough rationale for the approach selection. Demonstrates deep insights into theme extraction and categorisation.
  - Sentiment Analysis: Innovative preprocessing and feature extraction techniques are applied. The analysis of sentiment distribution is comprehensive and insightful.
  - Simple QA System: Implements a functioning QA system with creative approaches for keyword extraction and answer retrieval. Demonstrates excellent understanding of user queries and provides accurate, relevant answers.

- Good
  - Thematic Analysis: Applies topic modelling with some explanation of the chosen method. Themes are well-identified and categorised with minor oversights.
  - Sentiment Analysis: Adequate preprocessing and feature extraction methods are used. Sentiment distribution analysis has minor areas for improvement.
  - Simple QA System: The QA system functions, with some strategy for keyword extraction and answer retrieval. Shows a grasp of user queries with mostly relevant answers.

- Satisfactory
  - Thematic Analysis: Basic thematic analysis using frequent words with some justification for the approach. Themes are identified but may lack depth or clear categorisation.
  - Sentiment Analysis: The sentiment analysis is satisfactory with standard preprocessing methods. The analysis of sentiment distribution is adequate but lacks depth.
  - Simple QA System: The QA system is functional but relies on basic keyword matching with limited consideration for the complexity of user queries. Answers are somewhat relevant but may lack precision.

- Needs Improvement
  - Thematic Analysis: Attempts thematic analysis but with limited success. Justification for the chosen method is weak or absent, and theme identification is superficial.

  - Sentiment Analysis: The sentiment analysis is with minimal preprocessing. Sentiment distribution analysis is oversimplified and lacks critical insights.
  - Simple QA System: The QA system struggles to effectively match user queries with relevant answers. Keyword extraction and answer retrieval methods often result in irrelevant responses.

- Poor
  - Thematic Analysis: Lacks a coherent approach to thematic analysis. Themes are poorly identified with no clear rationale for the method used.
  - Sentiment Analysis: The sentiment analysis is with inadequate preprocessing and feature extraction. There is little to no analysis of sentiment distribution.
  - Simple QA System: The QA system fails to provide relevant answers to user queries. Lack of understanding of keyword extraction and answer retrieval techniques, resulting in a system that does not meet the task's objectives.

Slides - Content (8 marks)

- Excellent: The slides to stakeholders covers all key areas of the project and product, giving a comprehensive overview.
- Good: The slides to stakeholders covers only some key areas of the project and product, but not enough to give a comprehensive overview.
- Satisfactory: The slides to stakeholders provides a basic overview of the project and/or product.
- Needs Improvement: The slides to stakeholders provides a limited overview of the project and/or product, and misses key areas.
- Poor: The slides to stakeholders is unsuitable and requires significant improvement.

Slides - Professionalism (4 marks)

- Excellent: Excellent use of bullet points and structure, with minimal proofreading errors and the right amount of text. Visualisations and graphics are clear and professional.
- Good: Attempt at using bullet points and structure, with some visible proofreading errors. Visualisations and graphics are relatively suitable.
- Satisfactory: The text and structure were composed with basic consideration for the audience, such as limited use of bullet points or supporting text. Visualisations and graphics are basic or have not been considered for professional presentation.
- Needs Improvement: The text and structure could have used further refinement. Problems include proofreading errors and a visibly limited effort at professional presentation. Visualisations and graphics require further consideration.
- Poor: Significant problems with text and structure, including (but not limited to) proofreading errors and poor use of text. Visualisations and graphics are unprofessional.

Presentation (10 marks)

- Excellent: Professional group presentation that adheres to the time constraints, appearing well-rehearsed and planned. Explanations are clear and each member understands their part.
- Good: The presentation is within the time constraints, appearing to have some attempt rehearsal and planning. Attempts are made to explain all points (though some missing) and show professionalism, and each member contributes in some capacity.
- Satisfactory: The presentation is noticeably under or outside the time constraints, appearing to lack proper rehearsal and planning. Not all points are explained well. Professionalism and/or teamwork could be improved.
- Needs Improvement: The presentation would have benefitted from rehearsal and planning, with explanations, professionalism and/or teamwork the subject of improvement efforts. The time was not utilised effectively.
- Poor: The presentation has significant issues, including (but not limited to) unclear or limited explanations, not utilising the time well, and visibly appearing to lack rehearsal and planning.

Ethical Assessment (4 marks)

- Excellent: Discusses all relevant ethical issues outlined in the task (e.g. personal information, privacy laws, ethical considerations about the product, and a review plan) at a professional level in the slides and presentation.
- Good: Discusses some of the relevant ethical issues (e.g. personal information, privacy laws, ethical considerations about the product, and a review plan) at an adequate level in the slides and presentation.
- Satisfactory: Only partially discusses some of the relevant ethical issues (e.g. personal information, privacy laws, ethical considerations about the product, and a review plan) at a suitable level in the slides and/or presentation.
- Needs Improvement: Barely covers the relevant ethical issues (e.g. personal information, privacy laws, ethical considerations about the product, and a review plan) at an adequate level in the slides and/or presentation.
- Poor: Does not cover any of the relevant ethical issues (e.g. personal information, privacy laws, ethical considerations about the product, and a review plan) to sufficient depth in the slides or presentation.