|  |  |
| --- | --- |
| **Project** | **Text Mining and Language Processing Project** |
| **Team Size:** | 5-6 members per team |
| **Due Date and Deliverables** | 1. Week 5 – Sunday, 16th February (Project Proposal Report)- 5% 2. Week 13 – 15 minutes before class (Final Project Presentation Slides) – 15% 3. Week 14 – Wednesday, 23rg Apr, 11:59 PM (Final Project Report) – 15%   See Deliverables section for more details |
| **Percentage of Overall assessment:** 35% | |
| **Consultations** | Email instructor to schedule consultations. |

# **1. Project Description**

* The objective of the project is to apply the techniques that you have learned in this module to a real data set to perform some useful analyses (extracting high quality information from text). The analysis should be connected to the business problem (e.g. the banker needs to understand the sentiments of the customer on his new product and accordingly make decisions for improvements). You should be able to provide the recommendations to the business user based on the analysis (e.g. From the customer sentiments, you observe that the customer finds the product is only useful for certain age groups and you recommend the banker to change the product aspects to attract other age groups as well).
* You are free to choose either your existing data set or crawl new data set. Please choose the data set which consists of more dimensions. For e.g. political tweets with location, shares, likes, time, etc. Or hotel reviews with the product information, ratings, location, etc. Or Instagram messages with tags, likes, etc. Your data set should have more than 50K documents. The max I would suggest is 5M documents. E.g. Tweets should be in millions for developing a decent text mining application. Whereas wiki pages of movie stars can be around 50k to develop a decent application. If the downloaded dataset is too large, you can use only a subset of it (e.g. use only dataset for the past 3 years or only from Singapore, etc.).
* You are required to perform at least **two text mining tasks** and **project demo** on the data set you have chosen. The tasks should be on the same data set. A list of possible tasks is given in the sub-section. You need to clearly state what you plan to do in the project proposal. For the **proposal stage** please provide **three text mining** tasks (this helps us to choose the most practical tasks for actual project implementation). The instructor will give you feedback and approve the proposal before you start the project.
* Each text mining task requires;

1. Analysis of text challenges and solution approaches. E.g. Your data consists of spelling errors, and you propose to use Wordnet.
2. Text pre-processing and preparation of features according to your task. For e.g. you may not require to stem the text for sentiment analysis task and also prepare an additional features such as #of likes, retweets, etc., which can add more context and improve the performance of the model.
3. Preparation of at least 3 different solutions for each task. E.g. If you are doing information extraction, then you need to prepare at least 3 designs (Model 1 that uses HMM. Model 2 that uses neural network and Model 3 that uses Rule based).
4. Comparison of the models and evaluate their performance. You may need human judges, or you may already have gold truth or you need to prepare gold truth to achieve this. The performance can be compared both qualitatively and quantitatively.
5. Finally, each task should be demonstrated with an example to show how your task is useful in decisions by using which high-quality information. The objective is to demo each task with an example and show the outcomes of the application.
   1. You can create a web application to show the predictions on new data. E.g. when a new review is provided, what is the sentiment label for it. Or when a news article is provided, what are key topics and short summary for the articles.
   2. You can use Tableau for demonstrating the patterns/trends. For example, you can show the themes by age, gender or time, etc.,

# **2. Text Mining and Language Processing Tasks**

1. Document classification: For this task, you need a data set that already has labelled documents. First, define the labels you want to use to classify documents. This could be topic-based class labels, or this could be something else such as spam/non-spam. Next, train a classifier using labelled training documents. Finally, test the quality of the trained classifier by using the classifier to predict labels of unseen documents. To achieve this, you can first hold out a subset of your documents for testing purpose and exclude them from your training documents. Finally, analyse the outputs from the best model in the business context
2. Document clustering: For this task, you can use any data set. The goal here is to characterize the document set by grouping documents into clusters. You can play with different numbers of clusters. After clustering, you need to describe the meaning of each cluster of documents with the help of Python and other libraries. For example, you can use the most common words of each cluster to help you understand the common theme of that cluster. Finally, evaluate the models and analyse the outputs from the best model in the business context.
3. Topic analysis: For this task, you can use any data set. The goal is similar to document clustering, but instead of placing documents into “hard” clusters, topic analysis methods such as LDA find topics, which are represented by distributions of words. These topics give a summary of the major themes covered by the document set from which topics are extracted. Finally, evaluate the models and analyse the outputs from the best model in the business context.
4. Information extraction: For this task, you want to extract named entities from the data set you have chosen. You can perform some further analysis on the extracted named entities such as finding the most frequent named entities or finding pairs of named entities that co-occur frequently. You can create structured tables from the data and populate them. Finally, evaluate the models and analyse the outputs from the best model in the business context.
5. Sentiment analysis: For this task, your goal is to find positive and negative expressions from the data set you have chosen. Not every data set contains many sentiment expressions so chose this task only if it makes sense on your data set. Finally, evaluate the models and analyse the outputs from the best model in the business context.
6. You can propose any other text mining or NLP tasks and the corresponding details. Examples: Question answering, summarization, discourse analysis for relationship extraction, semantic analysis in social media data etc.,
7. Talk to LLM models like ChatGPT for more project ideas and please add references in your presentation slides/ project reports.

**3. Deliverable 1 – Proposal Report (5%)**

*3 page report +appendix*

Your report should include the following:

1. Brief introduction about the selected business domain and text mining related business problems.
2. Present the data statistics and 4-5 text challenges in the data.
3. Prepare at least two text mining tasks related to the business problem. For example: Topical analysis and summarisation.
4. Explain each task, its benefit to the business and your approach to it in brief. For e.g., we extract sentiments from the customer emails. We use dictionary approach, neural network approach and classification algorithm like SVM.
5. Explain the approaches for evaluating the models. Your data may already have gold truth or how will you create it? What metrics will you use? E.g. precision, average, perplexity, etc.,
6. Some ideas on how you will demonstrate the project in the end – briefly for each task.

# **4. Deliverable 2 – Final Project Presentation (15%)**

*13-min presentation*

1. Each group will give a 13-minute presentation in week 13. All members should give the presentation. All members should be prepared to answer questions from the instructor.
2. The presentation should include a brief introduction of the data set and the motivation for the analytics tasks, solution approaches, some details of how the analyses are done, evaluations, in particular any challenges faced, and some example output from the analyses. Limitations of your project. You need to also specify your analysis on why somethings didn’t work and what can be done to improve them. You can share 2-3 out of class learning that you implemented in the project. The remaining content will be in the report. Choose key examples for the presentation.

# **5. Deliverable 3 – Final Project Report(15%)**

*10 page report + appendix*

You are also required to submit a final project report (10 pages + appendix). We will post a template to e-learn. The report should include the following sections:

* Project title, Group information: group members and each member’s email address.
* Background and motivation: What is the business domain? What is the data set used? What tasks are performed? Why are these tasks useful?
* Methodology: How have you performed the tasks? For each task, which processing steps have you used? It might be helpful to draw flow charts to show the steps.
* Results/findings: What are the analysis results? This could include the classification accuracy, the summaries of document clusters, the proportions of positive/negative expressions and sample sentiment expressions etc.
* Discussions: What have you learned from this project? Are there limitations with the techniques you have used? How do you think the analyses could have been done better if you had access to more advanced techniques? Gap analysis for failures.

**Note: The final report counts towards 10% of the final grade.**

# **6. Other Grading Factors**

In addition, we will consider the following:

* Intra-team evaluation at the end of the term. Will only be used to penalise the students if teams rates them low in the team work aspects.
* Punctuality in submission

|  |  |
| --- | --- |
| within 1 hour | 10% marks deductions off the total marks you would have received |
| each subsequent hour | Penalty will double (i.e., 20%, 40%, 80% and finally 100%) |

* You are strongly encouraged to submit early

# **7. Resources**

* You can find datasets from the below web sites:
  + <https://www.kaggle.com/tags/text-mining>
  + [www.data.gov](http://www.data.gov)

OR crawl data legally for your project. E.g. LinkedIn has API, New articles websites are open.

* Some preferred projects
  + Healthcare projects- Clinical notes (<https://sites.google.com/view/mediqa2023/clinicalnlp-mediqa-chat-2023>), Wiki pages on Diabetes, Forums (https://forums.fast.ai/t/psychology-dataset/98739/2 )
  + Education - MOOC Discussion forums data, Curriculum datasets
  + Legal datasets - LegalNER Dataset, legal case reports, etc.
  + News articles – Sports, Business, Healthcare, etc.
* You can use generative AI like ChatGPT or search engines like Google to help to finding a project/ ideas/ mentorship [.](https://poe.com/BPASProjectBot) Please declare the usage of ChatGPT or other generative AI tools in your project and provide a summary of how the tools was used in your project. Quote the generated content and cite the reference materials in the appendix of the writeup.