Nova Information Management School
BSc in Data Science

Text Mining 2024-2025

Group Project: *"Solving the Hyderabadi Word Soup"*

Samosas vendor near the Charminar Mosque, symbol of Hyderabad

Statement of Work

Carina Albuquerque **Artur Varanda**

Table of Contents

1. Outline

11. Primary Objective

For the Text Mining 2024-2025 Group Project, **groups of students** will analyse a **dataset of 105 restaurants** in Hyderabad *(state of Telangana, India)* and a **dataset of 10 000** of their **Zomato reviews** to train text mining models able to provide **answers to a set of information requirements**, including the following **primary requirements**:

- *(Requirement **§3311**)* How well can we classify a restaurant's cuisine type using the content of their reviews as input? *(**Multilabel Classification**);*
- *(Requirement **§3312**)* How well can we predict a restaurant's Zomato score using the polarity of their reviews as input? *(**Sentiment Analysis**)*.

12. Output

The output of each group's work will consist of a written **project report** describing the group's results and a **Jupyter notebook** containing the code and data that generated the results, both delivered no later than **December 20, 2024**. Furthermore, groups must be ready to defend their findings and to clarify and demonstrate their methods during the **Project Defence**, scheduled for **January 15, 2025**.

2. Context

21. Project Impact

211. As per the Text Mining Course Unit File, this Group Project contributes **35% to the overall Text Mining grade**.
212. **Failing to deliver the Group Project and to defend its findings during the Project Defence directly leads to failing the Text Mining Course.**

22. Problem Description

221. Despite the use of likes, stars, scores, selectors, categories, classes, and a plethora of other attempts to quantify user experience using predefined labels, free text comments often capture meaning that goes beyond what can be garnered from such methods. Therefore, mining text comments to discover unquantifiable insights can add significant value to an organization's processes and products.
222. To illustrate the previous argument, consider restaurant scores: it is a well-known phenomenon of metrology that benchmark scores in most customer review systems for a number of areas tend to approach the maximum value [1] [2] *(think about the last time you saw a restaurant with more than 1 000 reviews and a score of 3.0 on Google Maps; would you consider it an average restaurant, or a very bad one, given its score?)*. Consequently, predefined scores lose meaning quickly, while text comments and reviews retain their usefulness (*"the food took too long to arrive!"*, *"the price-quality ratio was completely off!"*, *"I ate better in the army than I did here!"*).

223. Given the value of text mining at capturing "e-Word-of-Mouth" [3], assume that the Hyderabad Tourism Board has contracted your team to conduct an independent analysis of the city's restaurant landscape using a dataset of Zomato comments, and to use the analysis results to train models able to categorize (cuisine type) and classify (quality score) any new restaurants given their initial reviews.
224. The Hyderabad Tourism Board hopes to use the results and models to map the distribution of cuisine types by borough, to understand how dishes relate to cuisine types and to each other *(i.e., What sort of dishes are spoken about when discussing a successful North Indian restaurant? How often are Biryani and Hamburger part of the same menu?)*, and to detect any anomalously low scores and specific topics that might trigger health and safety inspections.

23. Datasets

231. Groups will use two datasets retrieved from Kaggle [4] to train the requested models and answer the information requirements:

2311. "**105_restaurants.csv**", a dataset of 105 restaurants comprised of the following features:

Feature Name	***Feature Description***
Name	*Name of the restaurant*
Links	*URL of the restaurant's Zomato page*
Cost	*Per person estimated cost of dining (Indian rupees)*
Collections	*Zomato collections to which the restaurant belongs to*
Cuisine	*Labels that describe the restaurant's cuisine type*
Timings	*Business hours and dates*

2312. "**10k_reviews.csv**", a dataset of 10 000 restaurant reviews comprised of the following features:

Feature Name	***Feature Description***
Restaurant	*Name of the restaurant being reviews*
Reviewer	*Person that wrote the review*
Review	*Text of the review*
Rating	*Customer rating (1 to 5)*
Metadata	*Reviewer metadata (number of reviews and followers)*
Time	*Time of the review*
Pictures	*Number of pictures posted with review*

232. Both datasets are connected one-to-many, so that every restaurant in the "105_restaurants.csv" was the object of several reviews in the "10k_reviews.csv".

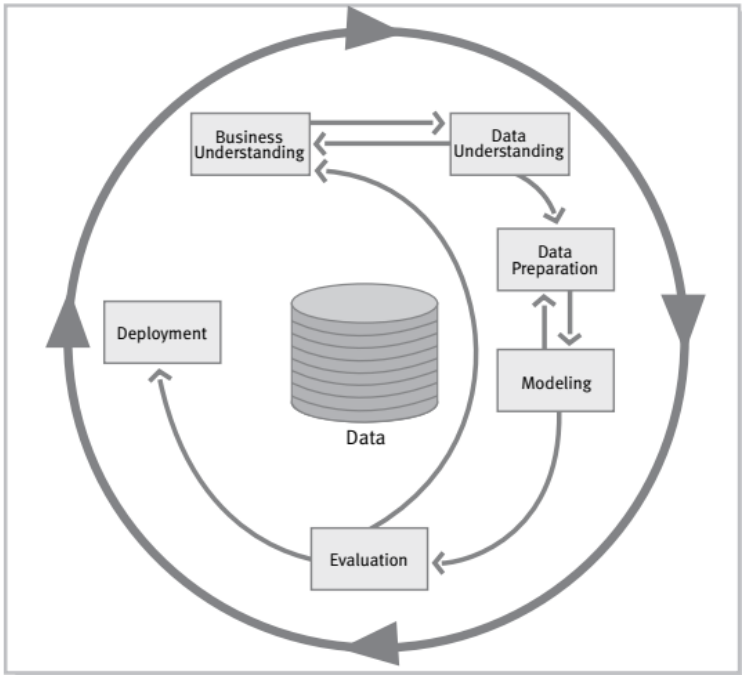
##Comment I: *You are allowed to consult the code available at the dataset's Kaggle page, but keep in mind that this project's information requirements are not necessarily the same as those of the notebooks, nor the methods therein are the best suited to solve the present project.*

24. The CRISP-DM Process Model

241. Given that this project can be thought of as a data mining problem [5], the Cross Industry Standard Process for Data Mining (**CRISP-DM**) [6] can be used as the project process model *(the sequence of steps that guides the project to a successful completion)*. CRISP-DM has often been described as the most popular analytics process standard [7].

242. CRISP-DM is a data-centric, four-level cycle comprised of six phases (first level) (**Figure A**), with each phase consisting of several generic tasks (second level). The phases and generic tasks are common to all data mining projects, so specific specialized tasks (third level) and processes (fourth level) must be defined for each project. The CRISP-DM manual is available at the **IBM website**.

Figure A – *The CRISP-DM phase (first level) cycle [5].*



243. For this project, groups are neither required to complete every second-level task, nor to explicitly define specific tasks or processes. However, the project tasks (**§32**), the project report template (**§34**), and the notebook template (**§3412)** follow the first-level phases, and groups are encouraged to complete the applicable second-level phases and to explicitly define and record (on the project report) some specific tasks and procedures – for example, a given group may create a "Data Cleaning" subsection in the "Data Preparation" section, and use it to state all preprocessing steps applied to the project datasets to prepare them for modelling.

##Comment II: *Note that the first phase of CRISP-DM, "Business Understanding" does not map well to the academic context of this project, so it is excluded from the notebook template (**§3412)**, and adapted to become a short literature review of similar problems on the report template (**§3411**).*

3. Execution

31. Intent

Our goal with the present group project is to allow the students to try their hand at performing an end-to-end (i.e., from raw data to actionable information) data mining process on a typical customer reviews dataset, not only to develop skills that are important in any workplace context (e.g., dealing with information requirements, writing reports), but also to highlight the importance of treating the various text mining methods as an integrated process (cleaning, feature extraction, modelling) whose various components must be purposefully selected – process design decisions matter as much as their implementation.

Consequently, we consider that your main concerns should be defining the appropriate pipelines (data preparation, modelling, and evaluation) for each information requirement and presenting the result in a clear, concise, and appealing report by leveraging the various visualisation methods.

We expect that by **December 20, 2024**, all groups will have successfully delivered their project report, published their code and data, and will be ready to present their findings during the Project Defence, scheduled for January 15, 2025.

32. Tasks

321. First, perform a short literature review that identifies similar or loosely related problems in peer-reviewed papers, focussing on the methods used to extract information and on performance benchmarks that you can use to assess your results. The literature review must contain at least two (02) and no more than five (05) references and should use the IEEE referencing system. *(Note that source trustworthiness will be taken into account, so citing Medium, TowardsDataScience and similar source should be avoided)*.
322. Then, explore both datasets – using visualisation techniques if necessary – to assess the amount of cleaning and integration required for the intended models.
323. Next, preprocess and integrate the two datasets so that *at least* the restaurant reviews are enriched with the cuisine types of the corresponding restaurant, or the restaurant dataset is enriched with the corresponding reviews.
324. After that, deploy and evaluate models to answer the information requirements listed below (**§33**).
325. Finally, write a report that summarises and illustrates your findings. The report must follow the template attached to this Statement of Work (**§3411**Erro! A origem da referência não foi encontrada.; **Annex A**), and must be *at most* ten (10) pages long, including annexes. **Preparing a clear, concise, and appealing report is your most important task.**

***##Comment III:** About report length: "I have already made this paper too long, for which I must crave pardon, not having now time to make it shorter." (Benjamin Franklin).*

33. Information Requirements

331. **Primary Requirements**

3311. How well can we classify a restaurant's cuisine type using the content of their reviews as input? *(**Multilabel Classification**)*.

3312. How well can we predict a restaurant's Zomato score using the polarity of their reviews as input? *(**Sentiment Analysis**)*.

332. **Secondary Requirements ("Targets of Opportunity")**

3321. What dishes are mentioned together in the reviews? Do they form clusters? Can you identify cuisine types based on those clusters? *(**Co-occurrence Analysis**; **Clustering**)*.

3322. Can the reviews be classified according to emergent topics? *(e.g., can review j be made up of 0.5 topic "service; speed; sympathy", and 0.3 topic "ambiance; decoration; furniture"?)* What do the emergent topic mean? *(i.e., are they meaningful regarding the project's context?)* Can relevant insights be extracted from the topics? (***Topic Modelling***).

34. Deliverables

341. To successfully complete the project, groups must deliver the following artifacts no later than **December 20, 2024**:

3411. A **Project Report** that follows the structure of the report template (**Annex A**), but not necessarily its formatting.

3412. A compressed (ZIP) folder comprising *at least* one **Jupyter Notebook** that follows the structure of the notebook template (**Annex B**) and the **data needed to run the notebook**. If using Python (PY) files to define helper functions or pipelines, the compressed folder must include them.

342. The **project report** and the **code/data compressed folder** must be delivered at the "202425 – Text Mining" Moodle page using the appropriate form.

343. The submission forms will remain open until **December 23, 2024**, but for each day of delay for any given deliverable, the final possible score will be decreased by one (01) full value. To illustrate, if a group delivers the Notebook on December 20, 2024, but only delivers the report on December 23, 2024, the maximum possible score for that group is 17 values; the inverse *(i.e., report on December 20, but notebook on December 23)* is also true (**§358**).

***##Comment IV:** About deliverables: "Finished is better than perfect." (Prof. Ricardo Santos, Nova IMS).*

35. Evaluation

351. The Group Project will be evaluated by comparing the project report against the **criteria defined in the scoring matrix** (**Annex C**), so that each criterion describes an idealized output *(sample criteria: "The introduction clearly restates the information requirements that the report answers")* against which the report is compared. Then, the evaluator assigns a number between zero (0) and one (1) to the comparison, with one meaning that the report under evaluation fully complies with the criterion, and zero meaning that the report does not comply at all. The compliance values are then multiplied by the corresponding weights, so that all weights sum to 20 values – the maximum possible score.
352. Criteria are grouped on sets of interrelated criteria. Sets are identified by multiples of 100, while subsets are identified by multiples of 10, so that a set (e.g., "800 – Cross-sectional") can contain multiple subsets (e.g., "810 – Formatting and Appearance"), with each subset containing specific criteria (e.g., "811 – The use of tables and figures is effective and increases comprehension"). The weights of every subset sum to form the subset score, while subsets sum to form the set score, and sets sum to form the overall score.
353. Apart from the set of criteria describing compliance with the notebook template (**Annex B**), the **code will not be evaluated directly**. Rather, wherever the report references a preprocessing, feature extraction, modelling, or evaluation technique, the code will be tested to ensure that the results referred to by the report are the result of the code and data submitted by the group.
354. For every instance of **partial compliance** with a criterion, the evaluator will request a clarification or demonstration to the group during the **Project Defence**, which can increase the compliance score up to 0.80 for that criterion for the entire group if the clarification or demonstration is effective at displaying mastery of the subject.
355. Additionally, the evaluator can request a clarification or demonstration from particular students during the **Project Defence**, and a successful answer can increase his/her final score. Conversely, failure to answer adequately a question posed during the Project Defence (even if the question is related to a fully compliant criterion) can decrease the final score for the student being asked the question.
356. Before the Project Defence, groups will be graded as a unit (e.g., every group member will receive the same grade), but individual scores are liable to change afterwards.

357. The full **Scoring Matrix (Annex C)** will be made available on Moodle on **December 27, 2024**. It will not be made available earlier to avoid guiding the groups' work directly to the expected solutions.

358. For each day of delay for any given deliverable, the final possible score will be decreased by one (01) full value.

359. For clarity, the overall distribution of the weights by sets and subsets of criteria are described on the table below, which summarizes the scoring matrix (**Annex C**):

Criteria ID	Criteria Set/Subset	Weight (out of 20)
100	Introduction	**1.50**
200	Literature Review	**1.00**
300	Data Understanding	**2.00**
400	Data Preparation	**4.00**
410	Data Cleaning	1.50
420	Data Preparation for Multilabel Classification	1.00
430	Data Preparation for Sentiment Analysis	1.00
440	Data Preparation for Co-occurrence Analysis	0.25
450	Data Preparation for Topic Modelling	0.25
500	Modelling	**5.00**
510	Multilabel Classification	1.50
520	Sentiment Analysis	1.50
530	Co-occurrence Analysis	1.00
540	Topic Modelling	1.00
600	Evaluation	**3.00**
610	Evaluation of Multilabel Classification	1.00
620	Evaluation of Sentiment Analysis	1.00
630	Evaluation of Co-occurrence Analysis	0.50
640	Evaluation of Topic Modelling	0.50
700	Conclusion	**1.50**
800	Cross-Sectional	**2.00**
810	Formatting and Appearance	1.00
820	Creativity, ambition, and flourish	1.00
		20.00

##Comment V: *About score: the Pareto principle states that for many outcomes, roughly 80% of consequences come from 20% of causes.*

36. Additional Instructions

361. **Admissible Methods ("Rules of Engagement")**

3611. For any given pipeline, pretrained word embedding models are allowed, but must not be the only feature extraction method (e.g., groups can use pretrained BERT[1] models for feature extraction but must necessarily compare their performance with a feature extraction method trained on the corpus – such as TF-IDF[2]).

3612. The use of pretrained models for topic classification is forbidden.

[1] Bidirectional Encoder Representations from Transformers

[2] Term Frequency-Inverse Document Frequency

3613. The use of pretrained models for sentiment analysis and named entity recognition is allowed.

362. **Reports**

3621. All bibliographical references must be made using the IEEE reference style. The style guide can be found at the following reference: [8].

3622. Reports cannot exceed ten (10) A4 pages, including annexes.

3623. The report and code will pass through a process of plagiarism and AI generation checking.

3624. The report will be the primary method of evaluating the group's work. When preparing it, remember that a reader should be able to understand the work without needing to check the notebook, so any methods, steps, or results (including visualisations) not mentioned in the report will not be evaluated.

3625. Long theoretical explanations of topics covered in class should be avoided. If other techniques or algorithms not taught during theoretical or practical classes are used, a theoretical explanation of the algorithm/technique should be provided in the appropriate section.

363. **Project Defence**

3631. Questions posed during the Project Defence can be either posed to the entire group, or to a specific student, but every group member should be able to answer any question regarding the project – not having worked on a specific phase or model is no excuse for not being able to answer, so failure will be grade accordingly.

364. **Timeline**

When?	***What?***	***Who?***	***Where?***
27SEP24	*Deadline for Group Enrolment*	*Groups*	*Moodle*
03OCT24	*Statement of Work and Annexes A and B available*	*Professors*	*Moodle*
12DEC24	*Project Support Practical Class*	*Professors*	*NOVA IMS (room 8)*
20DEC24	*Deadline for the Report and Code Submissions*	*Groups*	*Moodle*
23DEC24	*Deadline for the penalised Report and Code Submissions*	*Groups*	*Moodle*
27DEC24	*Scoring Matrix (Annex C) available*	*Professors*	*Moodle*
15JAN25	*Project Defence*	*Groups*	*NOVA IMS (room TBC)*
TBC	*Group Project Grades available*	*Professors*	*Moodle*

4. References

[1] Power Reviews, "Ratings & Reviews Benchmarks: Average Av. Rating," Power Reviews, 2024. [Online]. Available: https://www.powerreviews.com/average-rating/. [Accessed 25 09 2024].

[2] Z. Dobrijevic, "Revealed: Industry benchmarks for customer reviews in 2022," DAC Group, 10 05 2022. [Online]. Available: https://www.dacgroup.com/blog/revealed-industry-benchmarks-for-customer-reviews-in-2022/. [Accessed 25 09 2024].

[3] T. Hennig-Thurau, K. P. Gwinner, G. Walsh and D. D. Gremler, "Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet?," *Journal of Interactive Marketing,* vol. 18, no. 1, pp. 38-52, 2004.

[4] Chirag_ISB (username), "Zomato Restaurants Hyderabad," Kaggle, 08 06 2020. [Online]. Available: https://www.kaggle.com/datasets/batjoker/zomato-restaurants-hyderabad/data. [Accessed 25 09 2024].

[5] ORACLE Corporation, "What is Data Mining?," 2013. [Online]. Available: https://docs.oracle.com/cd/E11882_01/datamine.112/e16808/process.htm#DMCON111. [Accessed 12 05 2020].

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, CRISP-DM 1.0 - A step-by-step data mining guide, SPSS, 2000.

[7] M. S. Brown, "What IT Needs To Know About The Data Mining Process," 29 07 2015. [Online]. Available: https://www.forbes.com/sites/metabrown/2015/07/29/what-it-needs-to-know-about-the-data-mining-process. [Accessed 12 05 2020].

[8] Institute of Electrical and Electronics Engineers, "IEEE Reference Style Guide for Authors," 2023. [Online]. Available: https://journals.ieeeauthorcenter.ieee.org/wp-content/uploads/sites/7/IEEE_Reference_Guide.pdf. [Accessed 28 09 2024].

List of Annexes

A	Report Template	DOCX File
B	Notebook Template	IPYNB File
C	Scoring Matrix	XLSX File (Available 27DEC24)