NLP Project: Customer Review Analysis

1 Administration and Hosting Platform

The dataset together with a detailed description on the content and the prediction evaluation is given on https://www.kaggle.com/competitions/atml-unige-2024. In order to join it, you will have to:

1. Create an account on https://www.kaggle.com/. Please use a "Display name" that identifies you, for example your last name. When logged in, you can join the "competition" page on:

   https://www.kaggle.com/t/365160f53c8342c79cadc96bb5c79fe7

2. When all team members have joined the competition page, you can create a team.

The end-of-project deadline in on **20 May 2024 at 23:59 Swiss time**.

2 Dataset and Goals

You will be working with customer review data about various businesses in North America. The data are in particular composed of textual reviews from customers for given businesses. Each review is accompanied by the customer's satisfaction rating of the business and its services, ranging from 1 (bad) to 5 (good). You also have access to additional information about each customer, and about each business.

Task 1. Your first goal is to model and predict customer satisfaction ratings, given their textual review for a given business. You are provided with a "train" data set, containing a series of reviews **with** the corresponding rating, and a separate test data set, containing an additional series of reviews **without** the corresponding rating. Your goal is to make the **best rating predictions** possible **on the test set**. More details are available below and on the Kaggle website page above.

Task 2. A business wants to open a restaurant in the city of Philadelphia, and asks for your help to understand the regional market. In particular, the following points would be crucial in your analysis.

- 2.1 An insight on what restaurant consumers generally seem to like (for example in terms of food, service, location, etc...).
- 2.2 An analysis of the evolution of food trends in the area over time, in terms of consumer preferences. Do the preferences evolve over time, or do they seem stable?
- 2.3 Imagine you have to present your findings to the business owner and his investors. What advice would you give to the new business, based on your findings?

3 Prediction Evaluation for Task 1

The predictions for Task 1 are evaluated against the ground truth with the ordinal accuracy metric

$$\text{Ordinal Accuracy} = \frac{1}{n_t} \sum_{i=1}^{n_t} |\hat{y}_i - y_i| \,,$$

where $y_i \in \{1, 2, 3, 4, 5\}$ are the true customer ratings and $\hat{y}_i \in \{1, 2, 3, 4, 5\}$ are your submitted predictions, for each test observation $i = 1, \ldots, n_t$. This is an ordinal classification task where your predictions should be one of the possible ratings 1, 2, 3, 4 or 5. More details are given on the Kaggle website above. There are two different types of rankings on Kaggle for Task 1:

1. **Public Leaderboard:** Each day you may (not mandatory) submit up to two prediction files. These predictions are directly evaluated on 30% of the test data (always the same, randomly chosen beforehand). This score on the Public Leaderboard and gives you an indication of the accuracy of your prediction. The Public Leaderboard **does not count** for the final evaluation.

2. **Private Leaderboard:** Before the end of the Kaggle competition, you can choose one of your submissions to be the final team predictions. After the end of the competition, those final team predictions are evaluated on the 70% remaining test data, resulting in the Private Leaderboard. The final scores in the Private Leaderboard determine the accuracy of your predictions.

4 Rules

1. You can participate in teams of 4-5 students (4 teams in total). All team members must be registered for the course's examination.

2. Predictions should be based only on the provided training data, the provided additional information and insights from the Public Leaderboard.

3. You should use the python programming language. You are allowed to use any pre-built modules or packages in python, as long as they are explicit in your code.

4. No cheating of any kind.

5. You have to explain your main prediction approaches in detail in the final notebook in markdown cells.

5 Deliverables and grading

You should document your approaches and the corresponding code in two "IPython Notebooks", one for each Task. They are due on **20 May 2024 at 23:59 Swiss time**. The notebooks should be in `ipynb`-format, well-structured and well-documented. The notebook for Task 1 should in particular contain your python code that outputs exactly the predictions of Task 1 that you selected for the final Kaggle evaluation. Code that is not relevant to show in the main notebook, can be submitted in a separate python utility script. The notebooks can be submitted on the `Moodle` course website in the respective assignment module; only one member per team has to submit the notebooks; make sure to write the names or studen IDs of all team members in the notebooks. A possible structure for the notebook of Task 1 might for example be:

1. Introduction: Description of data set, imports, exploratory data analysis and feature engineering

2. Description of the best predictive model used, comparison of different methods, tuning parameters analysis, model selection approach

3. Best model diagnostics, final Kaggle prediction, conclusion

You can concentrate on the best model, but you should also compare the results (e.g., training, validation and test errors) to the other approaches. The notebook should contain plots to illustrate your findings.
Task 2 is more open in nature as there is no specific target. We don't expect you to analyse all aspects of the problem. You can decide yourself on which approaches, summary statistics or analysis procedures you focus on. Grading will be based on scientific correctness, originality and presentation.
There will be **oral presentations** during the last lecture on **21 May 2024 at 16:15** where each team will give a short presentation (15min) of their results, and the teaching staff will ask some group and personal questions. The analysis and code in the notebooks, together with the oral presentation will count for 40% of your final course grade.