Data Competition: Twitter Sentiment Analysis

1 The competition

The competition is hosted on https://www.kaggle.com/c/mlunige2021. In order to join the competition, you will have to:

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

   https://www.kaggle.com/t/5a6b286d329a4894837c0c28eedb859c

   Please do not share this link with anyone else outside of the course.
3. When all team members have joined the competition, you can create a team.
4. The competition ends on **May 30, 2021 at 23:59 Swiss time**.

2 Data set and Goal

In this competition, you will be working with data from the twitter social media (https://www.twitter.com). The data are mainly composed of so-called tweets, short messages posted by users on the website. Your goal is to model and predict the sentiment of each tweet (positive or negative). A tweet is positive or negative if it used to contain positive or negative smileys, respectively (for example ":)" and ":("). The smileys have been removed from the text of the tweets, and you will aim at predicting the corresponding sentiment based on the remaining text. You'll be provided with a "train" data set, containing a series of tweets together with the corresponding sentiment, and a test data set, containing an additional series of tweets without the corresponding sentiment. Your goal is to make the best sentiment predictions possible on the test set. More details are available on the kaggle website page above.
Be aware that the data set size is rather large. Especially in the exploratory phase you do not need to use the whole data. You should not wait a day for something to compile, if you can achieve similar results or conclusions on a sub-sample in much shorter time.

3 Prediction Evaluation

The framework and the evaluation metric of the data competition are explained on the kaggle website above. There are two different types of rankings on kaggle:

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 (randomly chosen beforehand). This score will be shown in the Public Leaderboard and gives you an indication of the accuracy of your prediction. Note that the 30% of the test data are always the same data points. The Public Leaderboard **does not count** for the final evaluation.

2. **Private Leaderboard:** Before the end of the competition, you can choose one of your submissions to be counted for the final evaluation. After the end of the competition, the predictions of this submission will be evaluated on the 70% remaining test data and this gives rise to the Private Leaderboard. The final scores in the Private Leaderboard is taken to determine the winners of the competition. You should therefore choose a prediction for the final score that performs best on new data. This might not necessarily be the same model than one with the best prediction on the Public Leaderboard.

4 Rules

1. You can participate in teams of 3 students.
2. No cheating of any kind.
3. Predictions should be made only based on the training data and information you obtain from the Public Leaderboard.
4. 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.
5. You have to explain your main prediction approaches in detail in the final notebook as markdowns.

5 Prize

The first 3 teams in the Private Leaderboard will receive bonus points for the course (which will have 100 points in total).

- 1st place: all team members receive 10 bonus points.
- 2nd place: all team members receive 6 bonus points.
- 3rd place: all team members receive 3 bonus points.

6 Deliverables and grading

You will document your different approaches to solving the data competition and code in an "IPython Notebook", which is due on **May 30, 2021 at 23:59 Swiss time**.
The main notebook should be in `ipynb`-format and contain your python code corresponding to the kaggle submission you selected for final evaluation. It should output the same exact csv file that you uploaded on kaggle, so make sure that your code is reproducible.
If you have too much code, you can submit additional notebooks containing details for your less important models. Make sure that all the code for your main approach, all of your analysis, and model comparisons (you can refer to the other notebooks's results), are in your main notebook.
The notebooks can be submitted on the `Moodle` course website where we will create respective assignment modules. Only one member per team has to submit the notebook, but make sure to mark the names of all team members clearly on it.
The main notebook should describe in detail the data set and how you approached the problem. A possible structure might be:

1. Introduction: Description of the data set, imports, notebook structure
2. Exploratory data analysis and feature engineering
3. Description of the best predictive model used, comparison of different methods, tuning parameters analysis, model selection approach
4. Best model diagnostics and final kaggle prediction
5. Conclusion

In the the description you should concentrate on the best model, but you should also compare the results (training, CV and test errors) for the other approaches. The notebook should contain a reasonable number of plots to illustrate your findings.
The analysis and code in the notebook(s) will count for 40% **of your final grade**.