40.016 The Analytics Edge

Data Competition

Stefano Galelli, Bikramjit Das

SUTD

Term 6, 2020

Challenge

Participatory monitoring: Data collection based on observations gathered by citizens on social media, such as Twitter or Facebook

Example: Weather agencies could use tweets to identify regions exposed to natural hazards

Challenge: Can we trust all tweets?

Problem description

Task: Develop an algorithm that determines, with the highest accuracy, what sort of weather a given set of tweets references. Specifically, the challenge is to determine whether each tweet has a *negative*, *neutral*, or *positive* sentiment.

Data:

- *train.csv*: 22,500 tweets with the corresponding classification
- *test.csv*: 7,500 tweets, with no labels

Accuracy metric: ratio between the number of correctly-classified samples and the total number of samples in the test dataset

Computing environment: R. You can use any package.

Kaggle

- The competition will be carried out on kaggle.com
- Results on the test dataset will be split into a *public* and *private leaderboard*
- The link to the competition will be released on December 3, at 5 pm. Please refer to the file *Data competition.pdf*

Kaggle

Rules

- **One account per team**: Name your account with the name of the team's representative, followed by the team's number, e.g., Walter White (Team 1) $\rightarrow$ *walter_white_team_1*
- Team mergers are not allowed
- You may submit a maximum of 2 entries per day
- You may select only 1 final submission for judging
- Only R is allowed (but any R package)
- Methods not covered in class (e.g., neural networks) are allowed

Schedule

Date	Event
December 3, 2020	Announcement of the Data Competition
December 3, 2020 (17.00)	Publication of problem details and competition rules + Release of the training dataset + Release of the test dataset
December 11, 2020 (23.59)	Last opportunity for submitting the results on Kaggle
December 13, 2020 (23.59)	Submission of reports, code, and peer evaluation form

Note:

- Use kaggle to download the data and upload your predictions
- Use eDimension for report, code, and peer evaluation

Evaluation

The Data Competition is worth a maximum of 38 points, distributed as follows:

- 15 points for the private leaderboard
- 10 points for the public leaderboard
- 13 points for the report and code authentication

Report

It is a short document (maximum of 4 pages, font size 12) containing:

- A high-level description of the approach developed
- A short description of the results
- A brief discussion on interpretability and limits of the approach
- (An executive summary is not needed)

Code authentication

- All dependencies and files needed to run your code must be available and listed in a *readme* file
- Your code must produce the same results as you uploaded on Kaggle