The Analytics Edge (Fall 2020) – Data competition

Detailed problem description and rules

Link to Kaggle competition:
`https://www.kaggle.com/t/43a8b353f753436dbc8d959bf544ad0f`

1 Introduction

Participatory monitoring is a novel form of data collection based on observations gathered by local residents. Such data collection process can rely on a variety of tools, ranging from amateur equipment to social media (e.g., Twitter, Facebook), on which people post comments and observations concerning various environmental and socio-economic processes. Participatory monitoring is quickly permeating multiple scientific domains as an alternative, or addition, to professional scientist-executed monitoring. During a flooding event, for example, agencies, utilities, and service providers could rely on both radar data and tweets to pinpoint the most affected areas. Are tweets reliable? How can one extract useful information from tweets? This is where data analytics can give us an edge.

In this project, you are tasked with the problem of inferring automatically the information contained in a large amount of tweets concerning the state of the weather. Specifically, your task is to develop an algorithm that determines—with the highest accuracy—what sort of weather the tweets reference. **The analysis must be carried out in the R computing environment. You can use any R package.**

2 Schedule

The schedule of events for this data competition is outlined in Table 1.

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

Table 1: Schedule of events.

Other info about the data competition will be published on Kaggle, which will act as a portal for downloading the data and uploading the predictions for the test dataset. Reports, code, and peer evaluation form should be submitted on eDimension through a dedicated link.

3 Problem description

As mentioned in Section 1, your task is to develop an algorithm that determines what sort of weather the tweets reference. Specifically, the challenge is to determine whether a tweet has a negative, neutral, or positive sentiment. The following datasets are provided:

- *train.csv*: 22,500 tweets with the corresponding classification / sentiment. The integers 1, 2, and 3 indicate negative, neutral, and positive sentiment, respectively.
- *test.csv*: 7,500 tweets. Naturally, this dataset has no labels. It will be used to quantify the performance of the algorithms.

The performance of the algorithms will be then evaluated based on their capability of classifying correctly the sentiment of each tweet in the test dataset. In particular, the evaluation will be based on the **accuracy metric**, defined as the ratio between the number of correctly-classified samples and the total number of samples. Kaggle will calculate the value of the accuracy on two subsets of the test dataset, named *public* and *private*. The results on the public dataset will be available during the competition (*public leaderboard*), while the results on the private one will be available at the end of the competition (*private leaderboard*).

4 Grading

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

- 15 points for the private leaderboard;
- 10 points for the public leaderboard;
- 13 points for the report (and code authentication). The report 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, and a brief discussion on interpretability and limits of the approach. An executive summary is not needed. The reports must be submitted using the eDimension submission inbox. **When submitting the report, please also upload a zipped folder containing the code you developed. 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.**

5 Rules

- **One Kaggle 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