Satalia

Data Science Challenge

The Dataset

The dataset for this challenge consists of thousands of statements (quotes) made by politicians and other public figures. The data includes various information on each statement and on the person who made it. The truthfulness of each statement was also evaluated by expert human editors, who then assigned an appropriate truthfulness label.

All the data is included in the provided **data.csv** file.The file includes one comma-separated line for each statement. The line includes the following attributes (columns) for each statement:

- label: a truthfulness label assigned by human annotators. Possible values (from most to least truthful) are:
    - true, mostly-true, half-true, barely-true, false, extremely-false.
- statement: the text of the actual statement that is being evaluated
- subjects: a list of topics related to the statement. These topics are separated by a ‘$’ sign. For example:
    - state-finances$taxes
- speaker_name: the name of the person who made the statement
- speaker_job: the job title of the person who made the statement
- speaker_state: the US state associated with the person who made statement
- speaker_affiliation: the political or professional affiliation of the person who made the statement. For example:
    - ‘democrat’, ‘republican’, ‘journalist’, etc.
- statement_context: the context in which the statement was made. For example:
    - ‘a radio interview’, ‘a press conference’, etc.

Challenge Deliverables

The first deliverable is a Python package that implements a binary truthfulness classifier for statements such as those found in data.csv. The package should be accompanied by brief instructions on how to install and use.

While it is up to you to decide how you want to structure your package, it should include at least the following three functions:

- **Function 1 [Training]:** Submit a dataset of statements with the same format as data.csv, for model training and model validation.
- **Function 2 [Prediction]:** Submit a json object that includes all the attributes for a single statement (e.g. statement text, subjects, speaker name, etc), except the truthfulness label. The function then returns a True/False prediction.
- **Function 3 [Modification]:** Submit a json object that includes all the attributes for a single statement. First, it calls Function 2 to compute a True/False prediction for this statement. It then tries to modify the statement in order to reverse the prediction (from True to False or from False to True). While any of the statement's attributes can be changed, the goal is for the modified statement to be as close to the original as possible.

The second deliverable is a short presentation (no more than 10 slides) that describes your approach and main findings.

Evaluation Criteria

Your work will be evaluated based on:

- The quality of the results returned by your solution (predictive power, ability to modify statements).
- The quality and efficiency of your Python code.

How to submit

You have 1 week to submit your work. You can email your deliverables back to the same email address from which you received the challenge. Please do ***not*** share the challenge or your code in a public repository such as PyPI, GitHub, BitBucket, or any other similar website.

This challenge is meant to demonstrate the way you approach the problem and lay the foundations for a solid solution. We don't expect you to deliver a perfect solution in 1 week :-) Do your best in the time that you have and feel free to contact us with questions at any point!