Machine Learning Task
Octo.AI

1 Task Definition

Given the news headline, predict the correct label for each headline. A label is defined in terms of news headline being sarcastic or not.

2 Data

2.1 Format

Each record consists of three attributes:
- `is_sarcastic` : 1 if the record is sarcastic otherwise 0
- `headline` : the headline of the news article
- `article_link` : link to the original news article. Useful in collecting supplementary data

Note : Since the news headlines are written by professionals in a formal manner, there are no spelling mistakes and informal usage. This reduces the sparsity and also increases the chance of finding pre-trained embeddings.

This dataset is of high-quality without noise.

2.2 Datasets

`train.json` : contains 24209 samples for training your model. You can use some part of it as dev set.

`test.json` : contains 2500 samples without the `is_sarcastic` label. Evaluation will be done on this dataset.

A record in train.json is below:

```
article_link:https://www.huffingtonpost.com/entry/versace-black-code_us_5861f
befe4b0de3a08f600d5
headline:former versace store clerk sues over secret 'black code' for
minority shoppers
is_sarcastic:0
```

3 Evaluation

The evaluation metric to be used is **average f1-score** over the 2 classes defined in section 1.

$$\textbf{average f1-score =} \; \frac{1}{n} \sum f1 - score$$

where n is total number of samples in test set.

Exact definition of f1-score is here : `https://en.wikipedia.org/wiki/F1_score`

4 Submission

The task is complete upon:

1. The complete **source code** and instructions to run the modules.
2. A CSV file containing prediction results
3. A brief write-up on all the approaches used in the submission.

Note: You are free to use any programming language for this task. You are also allowed to use any open-source library for sub-tasks such as data pre-processing, training and evaluating.

5 Final Note

In case of any clarification regarding the above task, feel free to drop a mail at **manishgui.garg@gmail.com**.