Take-home assignment

Sentiment Classification from scratch.

The basic idea of project is to prepare data for **text classification** and train a binary text classification model that distinguishes negative and positive **sentiment** data. A positive sentiment can be like the following “I love dasani water.” Or “I am the only one who likes Dasani water.” Or “what a beautiful bottle that is.” A negative sentiment can be like the following – “I hate Coca cola”. Or “why do you think I can tolerate another nasty Trump presidency?” OR “please save me from this torture.”

You will work with a dataset coming from social media using **python** language from which 2 datasets have been extracted – one for creating the annotations for your model and one to be used for projecting the model trained.

1. Do exploratory data analysis to understand what the data to be projected is about (“Data_to_project.xlsx”) using NLP techniques and statistics.
2. Manually annotate at least 100 positive and 100 negative examples from the given annotation data set (“Data_internship_test_anonymized.xlsx”) (there are more than hundred examples for each so stop when you get 100 of each).
3. Use these annotations to build a model which can be either be a logistic regression or even a neural network, whichever works best on your data and explain why you chose the model that you did in the end.
4. Evaluate the model that you have trained by using the metrics of your choice – either precision or recall or F1-score and explain why you chose them.
5. Project your model on whole dataset (“Data_to_project.xlsx”).
6. Explore the results and explain the classification model by choosing a few pertinent examples by showing us the score of your model on these examples and evaluating if this is the correct label that your model has predicted.
7. What possible improvements can you propose if you are not happy with the results of the model?
8. [Additional] could you also propose some data visualization method for visualizing the results?

For data visualization you may use matplotlib, plotly, bokeh, seaborn packages, etc…

Feel free to use any python libraries (pandas, numpy, scikit-learn, scipy, tensorflow…) or Deep Learning framework of your choice.

Output format:

- A link to a Git repository with all scripts and comments so that it’s easy to replicate

OR

- Well-organized Jupyter notebook with comments and sections and .zip archive with the data containing the projections of your model