EEP596 Winter 2022 || Mini Project 2
Twitter Sentiment Analysis

University of Washington, Seattle

March 6, 2022

Mini Project 2 Project Guidelines

- Please form a team of size 2 to work on the project. If for some reason you aren't able to find a team partner, you can work by yourself as well.
- Even though you are working as a team, please do mention each of your contributions to the project in your report.
- Each of the team members is expected to contribute to improving and implementing the machine learning models for the project. I.e. don't divide up work as one person pre-processing the data and the other person running machine learning models.
- Treat the mini-project as an opportunity to work on and hone your Machine Learning and probability skills. Done right, this project can be a line or two on your resume!
- Feel free to brainstorm high-level ideas on discord or in groups. However, the implementation details, data hacks and solution should be something you as a team have come up with.

Project Description

- **High-level** You will get to work on a Twitter Sentiment Analysis Data set and work on developing your amazing classifier to understand sentiments in tweets.
- **Kaggle Data set** Along with the training data made available on Kaggle, there will be another "evaluation data set" that you will get evaluated on. This mimicks a real-world setting where you don't get to "peek" into the data you get evaluated on or A/B tested on. Each Kaggle submission will give you a metric (e.g. Precision/Recall or F-score) to tell you how well you did on unseen data. You can make up to 7 Kaggle submissions per day (if needed) until your final submission (before the project is due). The Kaggle contest link will also be made available on Canvas.

Project Description

- **Sentiment Analysis:** Identifying sentiments in conversations is a challenging task for humans. Imagine how challenging it would be for machines? This is where you can put to use all that you have learned in the course, and also bring in your creativity in feature engineering/data processing/modeling hacks, etc to get a good enough model to help identify sentiments in tweets. In this project, the data set will have "targets" or "labels" that can have 3 options for each tweet - positive, negative or neutral sentiment. Even if a sentiment was predicted right for a tweet, the question is if the prediction of the model is also explainable to a lay-person. On the other hand, if the prediction was off - What factors contributed to the missing prediction and how can this be improved - This is also something you can work on as your improve your models. Most importantly, have fun through the project!

Project Evaluation

- **Code:** Showcase all of your code - Including pre-processing of data, post-processing/evaluation metrics, and machine learning models you implemented in a Jupyter notebook. Also showcase plots that shows how you evaluated models for over-fitting (e.g. The comparison of Training/Validation loss curves).
- **Models:** Implement at least 2 Machine learning models for your mini-project and try and optimize each of them. You can pick the best among them for your final Kaggle score evaluation. At least one ML model is a non-deep learning baseline and at least one is a deep learning model.
- **Insights:** What insights did you gain by working on this data set? What are the pros/cons of each ML model? Describe your thought process as well in the choices you made as you set up the machine learning pipeline. Also mention how you took care of over-fitting if any.

Project Evaluation

- **ML Pipeline:** Describe the components of the ML pipeline you set up for your project and any additional modules you added in to help optimize.
- **Evaluation Dimensions:** Evaluate your machine learning models along the following dimensions: Performance (mention metrics and values you obtained), scalability (if you had to scale your model to a million training data points, which of the models would you pick, what bottlenecks would you face and rough idea on how you would scale), interpretability/explainability (which of the models are more interpretable/explainable to a lay person or your company CEO) and computational cost. You can even make a table to summarize your insights or have a section in your report on this.

Project Evaluation

- **Kaggle Submission:** Each team also makes a Kaggle submission for a Kaggle contest set up for this specific class. This will make it exciting to see how you compare with other teams and get inspired to do better. The Kaggle submission makes up 20% of your project grade. There is a bonus 5% of project grade for top 10 team finishes in the Kaggle contest. However, your project grade heavily depends on your code submission, Kaggle submission, your insights (read insights bullet) and the evaluation dimensions mentioned earlier.
- **Due Date:** The due date for the project will be March 17th 2022.