CS57700 Project

Due date: The Day of the Final, Thursday May 6, 11:59pm. No Late Submissions.

1 Final Project

In this assignment, you will experiment with different techniques to improve your model performance in HW1. In addition to the labeled dataset, you will also have access to an unlabeled dataset of Congressional Tweets (`https://github.com/alexlitel/congresstweets`), which you can utilize using any of the various NLP techniques you have learned throughout the semester. *You are not required to use the unlabeled data.* We will also be releasing a relevant subset of this data shortly.

Using the dataset, you must complete the following things, with an end goal of improving performance over HW1:

1. Implement a **relevant** baseline. Any extensions you suggest should build on that baseline. You should use this baseline model to measure the improvement obtained by your extnesion.
2. Suggest a meaningful way of extending the baseline and test it.
   - Extensions can use the unlabeled extra data or not.
   - Simply using a library on the labeled data is not enough (i.e., run BERT on the data), you have to justify how you use the properties of the specific dataset. Some ideas:
     - Use the unlabeled data in some way (how you use it is your choice)
     - Use inference
     - Use meta-data/real world knowledge (knowledge base for example)
     - Learn and use representations and connect it to the decisions.
3. As part of your submission, you will submit a write-up on Gradescope as a PDF File. It should include the following things and be no longer than 4 pages:
   (a) Description of the baseline and a suggestion of different ways in which it can be extended.
   (b) Motivation and justification for the extension you will do (one of the ones you suggested in the previous section).
   (c) Describe the suggested experiments you did and how each of them would demonstrate improvement.
   (d) Show the results and comment on them, did it work/help? What can be learned from this? Any insights about the data?
   (e) Future work: If you had infinite time + computational resources data, what would you do next? How would that potentially help?
4. The other part of your submission will be the labels of the test set. In this case, please submit your predictions to turn-in in the same format as HW1 (so just fill in the labels and create a `test_proj.csv` file (in HW1 we used `test_lr.csv` and `test_nn.csv`instead.) Make sure the format is the same as HW1!

Remember that your submission must consist of two parts!

1. Write-up: 4 pages long submitted on Gradescope

2. Coding Part: All the code you used to run the experiments along with a file called `test_proj.csv` that has the same format of HW1 and has the predictions of your best model.

1.1 Submission Instructions:

Upload your project writeup (no more than 4 pages!) as a **typed pdf** in gradescope: `https://www.gradescope.com/courses/251477`. If you have not been added, you can register using the registration code GE6PV2.

1.1.1 Written Part

- For your pdf file, use the naming convention `username_project_submission.pdf`. For example, your TA with username *mehta52* would name his pdf file as `mehta52_project_submission.pdf`.

1.1.2 Coding Part

You need to submit all your codes via Turnin along with the final predicted labels of your best model. Log into `data.cs.purdue.edu` (physically go to the lab or use ssh remotely) and follow these steps:

- Place all the files in a folder named `username_proj#`. For example, your TA with username *mehta52* would name his folder for the Project as `mehta52_proj`. **This naming convention is important. If the folder is not named correctly, there's no way to identify whose submission is that. Hence, may result in no grading.**
- Change directory to outside of `username_proj#` folder (run `cd ..` from inside `username_proj#` folder)
- Execute the following command to turnin your code: `turnin -c cs577 -p finalProjSpring2021 username_hw#`
- To overwrite an old submission, simply execute this command again.
- To verify the contents of your submission, execute this command: `turnin -v -c cs577 -p finalProjSpring2021`.
  Do not forget the `-v` option, else your submission will be overwritten with an empty submission.