DSCI 552: MACHINE LEARNING FOR DATA SCIENCE

PROBLEM SET 5

Instructors: Dr. Kristina Lerman (lerman@isi.edu) and Dr. Keith Burghardt (keithab@isi.edu)

Deadline: Thursday, April 8, 2021, 10 A.M PDT

You can submit the report on Blackboard and code on GitHub Classroom. As long as the problem set is open, you will be able to upload multiple answers (the last attempt will be graded).

Task (20 points)

You are a new hire in a mid-size company. You have just completed your fourth task. You hoped to have a moment of rest however, the rest was not given to you. The CEO called you again. You enter the conference room. Moment later the technical manager and the developer enter the room as well.

The CEO said: "You did a solid job. We are happy to have you in our company. However, we have another task for you. We got a federal contract to develop a Crisis Management System. One of the components is supposed to monitor the epidemiological situation in the country. We have various teams working on this project. Each team works with different data sources - some public, some proprietary. We would like you to propose a model, that can be used to analyze Twitter messages.

You will get a collection of (labeled) tweets describing or commenting the local Covid situation. We would like you to build a classifier, that can sort the messages into 5 categories: Extremely Negative (0), Negative (1), Neutral (2), Positive (3), Extremely Positive (4). When we have it, later, we will be able to build a system that measure how the perception of the epidemiological situation change in various regions - and this can be helpful in planning the next moves by various federal agencies.

I would like to see a preliminary report in two weeks. We will meet again on Thursday, April 8, at 10 am."

Hint: This task is to build a text classifier. You have to first tokenize all words. There are various libraries that offer Tokenizers, e.g., NLTK (Natural Language Toolkit). TensorFlow have also a tokenizer, that you can use. You can also test, if stemming (read about it) can help you.

Your Technical Manager said: "You have several choices how to approach this problem. You can use a Naive Bayes Classifier for start. Or you can train your own word embedding. If you feel like you can, you can also try recurrent models or you can use transfer learning techniques, by taking a pre-trained word embedding. You can also try to clean the text data by e.g., stemming the words."

Hint: You do not have to use transfer learning techniques. This is just an optional extension. The same about recurrent models. It is just an option for people who would like to do a little more. As always, you should validate your model choice - by for example, comparing the performance of some candidate-models on some validation tests.

The senior developer took you aside and said: "My task will be to maintain your code. Please, write comments and try to explain any nontrivial part of your code!"

Data

There are two data files. The first file, ps5_tweets_text.csv, contains labeled Tweet messages. For your convenience, I prepared the labels in two formats: text and numeric, see ps5_tweets_labels.csv and ps5_tweets_labels_as_numbers.csv.

Report

To help you, I prepared a template. See https://www.overleaf.com/read/vnvhqxkpdhbk. You are encouraged to use the template, but you are free to use other editors or make modifications. Just ensure that the final submission of the report has to be in PDF. Submission of report has to be done on Blackboard. As always submit your code on GitHub classroom and report on Blackboard.

Code Submission

We have created a GitHub Classroom where you can create private repositories. We will update the class on Piazza on how to go about uploading your solutions on this platform.

Grading Rules

In order to grade your work, we will role-play the following situation. We will assume, that you are a new-hire in our company. You are asked to provide a comprehensive technical report that illustrates your findings. We will evaluate it from the perspective of three people.

- Your CEO (she would like to hear high level stuff. She will probably only read the conclusions and look at main figure). (4 points for report).

- Your manager (he would like to see a detailed report; he might also look at some parts of the code). (6 points for report and 2 points for code).

- A senior developer (they would like to see the code and won't read the report at all). (8 points for code).

Your final score is: 10 points for report and 10 points for code.

Don't Panic

Don't panic. We understand that this is a large, open ended task. We also understand that this might be the very first technical report that you were asked to write. We are dedicated to help you do your best work all while keeping the standards high. We acknowledge that you have limited time and resources to complete the task. This report doesn't have to be perfect for 100% score. If something is not clear, ask your questions on Piazza or reach out to us during office hours.

Note: Cite any source you use (even if you adopt/copy a snippet of code). Failure to do so would amount to plagiarism.