Kaggle Competition: Congressional Tweets

Please read all of the guidelines carefully before submitting the project.
☺ There are ***100 points*** in total. **You can work alone or in a group of two in this project.**

Due date: April 8th, 2022 (Friday), 11:59 PM. Late submissions will be penalized as per syllabus by 10 points / day. No submissions will be accepted two days after the deadline.

Deliverables:

1) The code of the project in **.ipynb** format (one file)
2) The project report written in ***IEEE format*** with ***LaTeX*** and exported in **.pdf** format (one file)

Guidelines – Before You Start

1) You will be using the **`Python`** programming language for this project. You need to write your codes in an empty **.ipynb** file.
2) Make sure that you provide many comments to describe your code and the variables that you created.
3) Please use the ***IEEE*** journal template on ***overleaf.com***. Here is the link: https://www.overleaf.com/latex/templates/preparation-of-papers-for-ieee-sponsored-conferences-and-symposia/zfnqfzzzxghk
   To be able to work on ***overleaf.com***, you will need to register first (you can also compile your ***LaTeX*** file locally.)
4) For some of the code, you may need to do a little bit of "*Googling*" or review the documentation.

What is Congressional Twitter data?

This dataset consists of all tweets extracted from all Congressional politicians who use Twitter. The data covers a period between the years 2008 - 2021 and contains 857,803 tweets. You are provided with three datasets:

- **Training data**: 592,803 tweets
- **Test data**: 265,000 tweets
- **Sample submission data**: A concise version of the test dataset that includes only the `Id` and `party` columns

Data Dictionary

Features:
Id : id number associated with the tweet
favorite_count : number of times the tweet was favorited
full_text : full text of the tweet
hashtags : list of hashtags included in the tweet
retweet_count : number of times the tweet has been retweeted
year : year of the tweet

Class label:
party : party of the owner of the tweet [**D** = 'Democrat', **R** = 'Republican']

Assignment Goals

In this assignment, you are expected to achieve two primary goals:

1) Correctly predicting the ***party*** of the owner of tweets (**60%** of your grade)
2) Writing a report describing the data and your methods (**40%** of your grade)

Making Predictions (60 points)

Idea: For the Kaggle competition, you are asked to use the features described in the ***Data Dictionary*** section above to predict the ***party*** variable [this is a two-class classification problem]. For your predictions, you are welcome to use any existing method (including the methods we have learned in class and beyond.)

Resources and Suggestions: To tackle the classification problem in this assignment, you will need to use some classification techniques that are frequently employed in the context of NLP. Here is a good place to start reading: https://www.nltk.org/book/ch06.html.

Access to Datasets, Signing Up, and Submissions: To sign up for the competition, to download the datasets that you will need train your model, and to submit your predictions, please check the following page:
https://www.kaggle.com/c/congressionaltweetcompetitionspring2022

Preparing the Submission Dataset: To prepare the submission file that you will need to submit to be ranked on `Kaggle`, you will need to update the ***sample_submission.csv*** file provided. The `Id` variables in the **test data** and the **sample submission file** match and they represent the same tweet.

Submissions and Evaluations: You can send up to **20** submissions per day. Please use your **real name(s)** on Kaggle so that we can track your ranking. You can form groups of two on Kaggle. Your submissions will be evaluated by using the ***accuracy*** metric that is calculated from a sample of observations coming from your submission file. This will allow us to rank the strength of your model among your peers.

Grading for Kaggle rankings: There will be two separate lists of ranking for undergraduate and graduate students (that we will internally create). If your team has one undergraduate and one graduate student, you will be placed in the graduate student ranking list. Your grade will be determined by your ranking.

Academic Honesty: Please work separately from other teams and concentrate on your project only. Your code will be evaluated by using an automated procedure to make sure that the participants have followed the rules regarding academic honesty. Code sharing with other teams is not allowed.

Writing the Report (40 points)

In the second part of your assignment, you will be writing a short report describing the dataset (i), and describing the methods you used to create your classifications (ii). Please follow the following guidelines:

Length

- If you are a team of ***undergraduate*** students, please write a report of **2** pages
- If you are a team of ***graduate*** students, please write a slightly more detailed report of **3** pages

Visuals will be regarded as part of your report, references won't. Report will be in ***IEEE*** format.

Structure

Your report needs to have the following four sections:

1) ***Introduction***: A summary of what you expected and did, and two-three of your most significant findings (please use some numerical results here)

2) ***Descriptive Analysis***: Introduce your descriptive findings about the dataset here

3) ***Classification Methods***: Provide a description of your strategy and the steps you took to improve your classification model (this includes the steps you followed for data-preprocessing, setting up the model, and checking the strength of the model)

4) ***Classification Results***: A detailed discussion on the results you obtained. What is your accuracy score? Evaluate and criticize yourself / your team.

Descriptive Analysis

You are free to explore different methods of descriptive analysis to help the reader understand the data better. A few ideas to consider (not an exhaustive list):

- Sentence length, sentence structures, word choices, frequency of words
- Tweet patterns comparing Republicans and Democrats
- Analysis of temporal trends
- Use of hashtags, frequencies, meanings
- Topic analyses (using different methods)
- Clustering and dimension reduction

If you are a team of ***undergraduate*** students, please include **4-5** visuals and tables in your report.
If you are a team of ***graduate*** students, please include **6-7** visuals and tables in your report.

Grading of the Report

You will be assigned a numeric score (out of **10**) for each of the sections [detailed in ***Structure***] that will need to be included in your report. The score for each section will be based on:

- Completeness
- Level of attention to detail
- Quality of the work produced
- Complexity of the task(s) and method(s) chosen
- Multitude of ideas and creativity
- Adherence to guidelines (such as length and number of visuals/tables)

***

Final task: Place all of your code in one file and save it as **congressional_tweets_project_code.ipynb**.
Save the project report as **congressional_tweets_project_report.pdf**.