**NLP Project Proposal**

Authors: Nigel Martis, Rohan Paul, Bradley Reardon, Jongchan Kim

• What problem did you select and why did you select it?

The problem we selected for this project is voter turnout in American elections. There was a lot of speculation around voter turnout in 2020 with both high-profile candidates and new voting methods due to the pandemic. Furthermore, there is a long-standing interest in why many eligible voters decide not to vote. Turnout has been around 55-60 percent in most modern presidential elections, but rose to 67 percent in the highly salient 2020 presidential election, which was the highest turnout since 1900. However, that leaves around one-third of eligible voters still deciding not to cast a ballot. We want to explore reasons that they may not have voted.

• What database/dataset will you use?

The dataset that we will use is sourced via Stanford University and a paper written by Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei titled *3D Object Representation for Fine-Grained Categorization*. It contains 16,185 images of 196 classes of cars and is split into 8,144 training images and 8,041 testing images which each class being split roughly 50-50 between the two subsets. Classes are typically at the level of *Make, Model, Year*, e.g. 2012 Tesla Model S or 2012 BMW M3 coupe. This dataset is large enough to train a deep network.

• What NLP methods will you pick from the concept list? Will it be a classical model or will you have to customize it?

We plan to use the <insert pretrained model name> while using our dataset to fine tune the model to allow for more precise classification of images of vehicles by *Make, Model, Year.* We will adjust the model if a need arises while testing the model.

• What packages are you planning to use? Why?

The packages we will use are NumPy, Pandas, Matplotlib, and PyTorch. NumPy and Pandas will be used for data cleaning and preprocessing, Matplotlib will be used for visualizing our findings, PyTorch used to implement, train, and test our models.

• What NLP tasks will you work on?

The reference materials we will use to obtain sufficient background on applying the chosen network to our vehicle image classification are the NLP course materials, the official documentation websites for the packages we will be using, and various websites such as medium.com.

• How will you judge the performance of the network? What metrics will you use?

We will judge the performance of our network by checking the accuracy score of our test dataset. The metrics we will calculate are precision, recall, F1, and MSE for loss. We may use accuracy as a metric, dependent on how balanced the dataset is. This will be examined during exploratory data analysis.

• Provide a rough schedule for completing the project

The table below outlines the progress and deliverables that we plan to achieve at each date.

|  |  |
| --- | --- |
| Proposed Date of Completion | Actions/Deliverables |
| Tuesday, March 30 | * Decide on a dataset * Create GitHub repository |
| Tuesday, April 6 | * Submit group proposal * Begin exploratory data analysis |
| Tuesday, April 13 | * Clean and process the data for train-test into random forest and logistic regression models * Apply the models to the data |
| Tuesday, April 20 | * Evaluate the results of the models * Refine models if needed * Begin writing the final report |
| Tuesday, April 27 | * Create GUI * Create presentation |
| Sunday, May 2 | * Recording of presentation * Finalize final report * Individual final reports |
| Monday, May 3 | * Final Project |