CAPSTONE PROJECT PROPOSAL: A STUDY ON ECHO CHAMBERS

By: Diego Cuevas Gonzalez

Goal of the project:

The goal of this project will be to establish a relationship between political YouTube content and the type of engagement that it receives. It will basically be a social study on if YouTube is prone to the creation of echo chambers, which is "an environment where a person only encounters information or opinions that reflect and reinforce their own." If an outstanding majority of the comments under a video are in line with the opinions of the video itself, I will consider that YouTube is indeed prone to the creation of echo chambers.

Data collection:

I will use a mix of YouTube’s API and Beautiful Soup to extract the subtitles of several videos as well as comments. Since I’m planning to use word embedding, which pre-trained models may not include names, I will avoid videos on political figures and focus on broader subjects. Furthermore, for the sake of fairness, I will include equal amounts of videos of each subject from each side of the political spectrum. To add a layer of complexity to this project, I could potentially also include the number of likes under each comment and/or under the video itself.

Algorithms used:

Since I will be dealing with text, I will be using word embedding algorithms to make the collected data interpretable for the models used in the following steps. Afterwards, for better accuracy, I will use a supervised learning model classifier. I will manually assign the value of 0 to text that comes from a liberal video and 1 to text that comes from a conservative video. I will use this data to either train a logistic regression model or a Neural Network model for binary classification. Once I obtain a model with satisfactory accuracy, I will use it to determine whether the collected comments are liberal or conservative. Finally, I will use a linear regression model to determine if YouTube videos are prone to inhabit echo chambers based on the subject and the political affiliation.

Extra key points:

* Use scripted videos, not debates, to reduce noise (redundant words)

Notes:

I Initially intended to include one video per bias (left, leaning left, center, leaning right and right), per topic (abortion, gun control and affirmative action). However, during the data collection, I saw that some of the news medias that AllBias.com categorized as leaning left and leaning right did not have a YouTube channel. Those who did also lacked severely in engagement, having often less than a thousand comments in the videos of interest. I therefore decided to avoid using those categories in my studies. Instead, I collected data from at least six videos per remaining bias, all from different news sources, per topic.

Since my model needs to be trained on labeled data, it will mostly be trained with transcripts. However, to vary the length of the strings fed into it so that it is more accurate in classifying comments, which is its final goal, I will also feed it transcripts of shorter videos. In this case, I will not be pulling the comments of these videos. Therefore, engagement is irrelevant for these videos.