Math 189r: Mathematics of Big Data, Final Project Proposal

Justin Jiang, Andy Liu

June 15, 2020

For our final project, we propose using Natural Language Processing (NLP) and text sentiment analysis to compare public sentiment with media sentiment in news relating to the 2020 Presidential Election, specifically regarding Democratic Nominee Joe Biden and current President Donald Trump. We plan on using publicly available curated data sets, such as the Reuters-21578 benchmark corpus. We also plan on scraping both online news sources and public forums such as Reddit in order to collect data. We will specifically target headlines or posts connecting one or both of the major party candidates to specific issues, which we hope to use to estimate public sentiment and news/media sentiment towards the candidates on specific issues.

Then, we will use Python libraries such as scikit-learn and the Natural Language Processing Toolkit to analyze our data. By programming a web crawler, we hope to generate a live data feed that we will feed into our machine learning model to continuously inform our model on each candidate and topic. In order to reflect the wide variety of topics that the candidates will take stances on, as well as the wide range of public opinions that we will be able to analyze, we will be using Aspect-Based Sentiment Analysis. This will entail analyzing our corpus with a topic analysis model and a text sentiment model before combining our outputs to find sentiment on specific topics.

In the current political atmosphere, many say that the news media is often disconnected with the general public. We intend to the veracity of that belief by using methods to estimate the sentiment both the public and the media hold against controversial topics important in the 2020 Presidential Election. For example, say we were to estimate sentiment on gun control, a longstanding partisan issue. If we analyze sources that are representative of both the public and the media, then we can estimate the sentiment that each group holds towards gun control. Say we find that the media is overwhelmingly negative on gun control (i.e, supporting widespread gun reform), but we also find that the general public is actually more moderate (supporting some gun reform, but not too much), then we can say that there is a gap between media and the general public. We can scale this method to important topics in the upcoming election to calculate a final "distance" score of how separated the media is from the public.