CSE 5544: Project Proposal

Ron Davies & Logan Frank
11/22/2020

Title: Visualizing the U.S. Presidential Election

Introduction: The United States (U.S.) presidential election (including the debates and months leading up to the election) is a major event that occurs every 4 years. Arguably, this year was one of the most important and influential elections in recent history. In each election, there is ample amounts of data produced from the debates, marketing campaigns, Twitter, Google searches, and more. With so much data being produced from these sources, it is hard to ingest all of this data and no adequate visualization dashboard exists for ingesting it. It is important that we be able to visualize all of this data and it can often reveal insights about the candidates that the media does not choose to report.

Thus, we aim to create a web-based dashboard that presents factual information from a variety of data sources through easy-to-understand visualizations of election-related data. The time range of data we plan to visualize will start with the day the primary candidates for the Democratic and Republican parties were decided until 11/17/2020 (2 weeks after Election Day). Although the 2020 election has already been decided, the dashboard we plan to create could be easily modified to work using streaming data for the next presidential election, providing real-time information as data is updated and provided and giving voters information beyond media outlets to help them make an educated decision on who they will vote for. High-level questions we plan to ask are:

- What is being discussed?
- How are people reacting to the conversation?
- Is there a high or low sense of professionalism conveyed audibly?
- Are there general trends that can be seen over the course of the election season? Particularly regarding political issues such as environment, COVID-19, gun control, etc.
- How do major events influence an election?
- Are there trends occurring in other data sources that can be seen over the time frame of a presidential debate?
- What can be visualized about a candidate's history to give insights about their career?
- How do other countries perceive the U.S. election?

Detailed Analytical Questions:

- What major topics are being touched on in the most recent 3-5 minutes?
- Is it easy to distinguish one voice from another during the current segment of the audio?
- What are the current suspected swing states?

- What political topics are trending in these swing states?
- How does a candidate's word cloud change from topic to topic during a debate?
- How does the type of dialogue differ between the candidates?
- How do Google search trends change throughout 2020? Particularly regarding political topics.
- What are other countries talking about (on Twitter) during the debates?

Datasets: We plan to incorporate numerous datasets into our visualization dashboard, with our main dataset being the "US Election 2020 - Presidential Debates" dataset from Kaggle. We will also use data from other sources such as Google Trends, Twitter, Kaggle, and more. We plan to align and combine these different sources in our visualization dashboard. The datasets we currently plan to use are:

- Kaggle US Election 2020 - Presidential Debates
- Google Trends (we will search different keywords and download their respective csv file)
- Twitter data using the Twitter API
- Kaggle US Election 2020 Voting Data
- #Election2020 Dataset

Visualization Requirements:

- Spectrogram: Frequently during the most recent debates, there were many interruptions and disruptions by both speakers. Providing a spectrogram figure will allow a debate viewer to observe the way the interruption may interfere with the audience's ability to perceive what was being said. It will also provide debate viewers with another metric by which to judge the pacing of each speaker. This may become most interesting to any viewer when a candidate is presented with a much more challenging question versus an easier one. By providing a small window of time to be visible as the debate progresses, this figure will enable any active viewer to draw their own conclusions on the candidate's responses using this data as justification.
- Word cloud: As different topics are covered during a debate, the most frequent words a candidate uses can help determine what their focus is for the given topic. Visualizing each of these word clouds can help voters understand if a candidate's priority is on addressing the topic and if so, what exactly about the topic, or if the candidate is attempting to avoid the prompt.
- Top Tweets: Popular tweets from swing states provide insight on the current influence of the debate responses from each candidate. Any viewers that have already chosen a side or are deeply curious about how the election may be progressing in different parts of the country will find a sampling of responses that represent portions of crucial voting communities. This may provide researchers with insight into social networks they should investigate or weaknesses in the latest polls. For causal observers, it provides a sense of community by presenting what other people think in regions that are highly contested rather than from their own inner circles.

- Language: Understanding the type of words a candidate frequents to using, how fast a candidate speaks, or how many false claims a candidate makes can tell a lot about their character, confidence, and trustworthiness. Being able to visualize and understand how a candidate discusses political matters is another important factor voters should consider when determining their election, this sort of visualization could aid in helping the voter make that choice.
- Other: It would be interesting to include some sort of high-dimensional data to where we could run a dimensionality reduction algorithm to visualize the data. What data this would be is yet to be determined.

Software: We will be using exclusively Python with the Dash package for creating our dashboard and Plotly to create the visualizations used.

Evaluation: We will evaluate our visualization system on the basis of figure clarity and interpretability. These aspects will be assessed by matching components of our figures to objective principles covered during this course in addition to informal survey responses. To get an idea of how others handle our visualization, we will take a short survey from a few peers. We will seek feedback on whether the spectrogram and its purpose are clear as that is one of our most complex figures. The feedback for the other figures and the dashboard as a whole will provide us with an indication of the overall effectiveness of our visualization system.

Team Members and Work Distribution: The data collection and cleaning will be done on an as needed basis by both team members in order to suit the tasks that each will be working on. Creating the visualizations will ultimately be divided up evenly; whoever wants to create what visualization can choose to do so. Tentatively, the following tasks are split up and assigned:

- Ron Davies: Spectrogram, trending Twitter posts
- Logan Frank: Word clouds, debate dialogue, other Twitter and Google search data