Visual Analysis of Newspaper Propaganda Biases

CSE 578 Data Visualization

I. INTRODUCTION

For my MS CS portfolio, I chose to work on a mentored project, with the support of our mentor Aditi Mishra. As a final project for CSE 578 Data Visualization, I worked on this with a group of six people. When Dr. Bryan suggested this topic, we had no idea what we were getting ourselves, but it piqued our interest. During the last decade, news articles have propagated a lot of misinformation about major events that have gotten a lot of attention, such as the 2016 US Presidential Election, Brexit, and the COVID-19 infodemic, to name a few. Despite increasing study on fact-checking and deception detection, little attention has been paid to the precise rhetorical and psychological methods used to express propaganda themes. [1] The main goal of this project is to raise public awareness through a visual analysis system about the propaganda strategies in order to promote media literacy and critical thinking in those who read the news, as well as to reduce the impact of "fake news" and disinformation campaigns. We decided to extend the existing PRTA (Propaganda Persuasion Techniques Analyzer) API, which allows users to analyze and compare articles crawled on a regular basis by emphasizing the spans in which propaganda techniques occur. [2]

II. DESCRIPTION OF OUR SOLUTION

Using the PRTA API, our system crawls many articles from a subset of the Kaggle "All the news" dataset [3] based on a user's search criteria and finds the percentage scores of various propaganda techniques used in each article. The user is then presented with this information via five visualization panels created by us that compare different publications based on the propaganda techniques they employ.

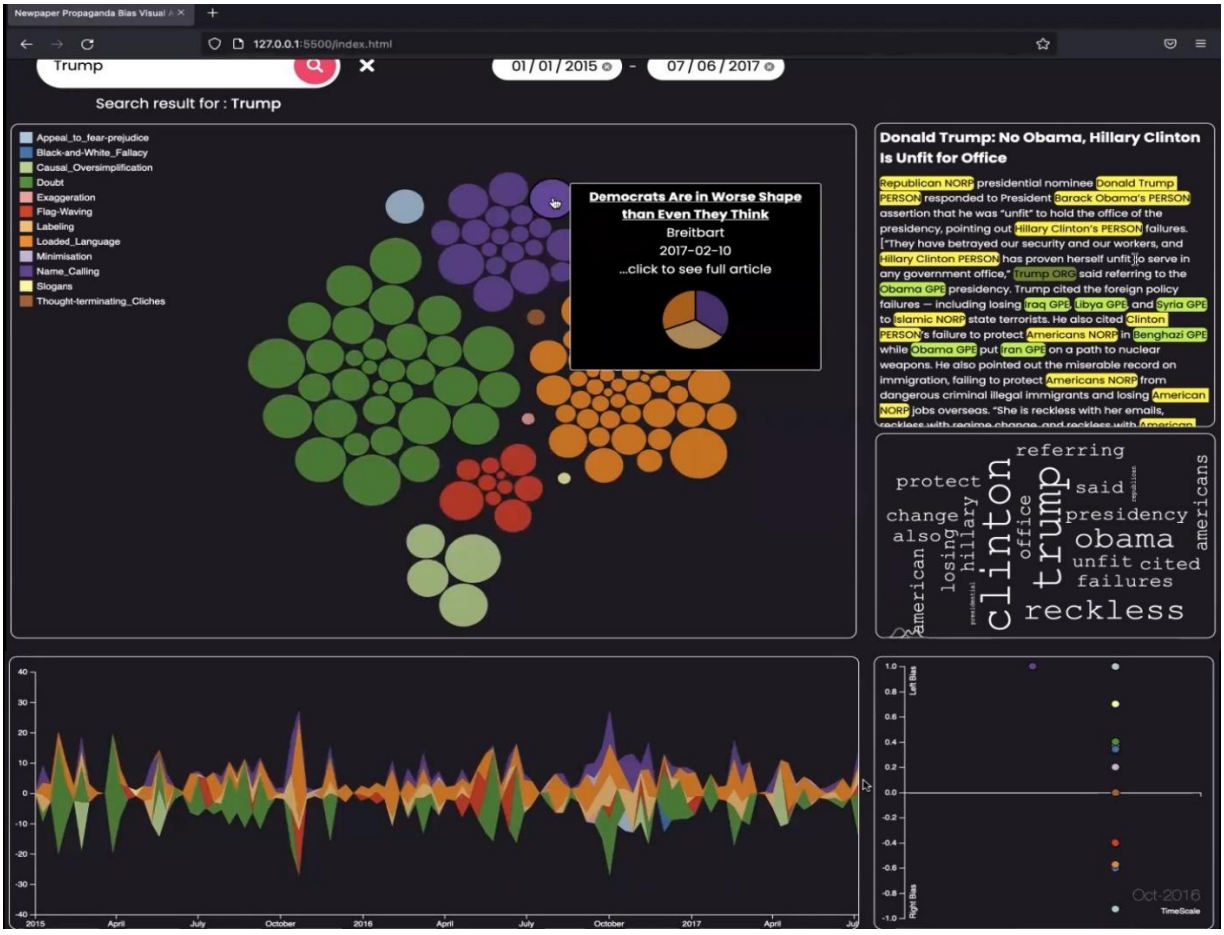

Fig 1: The Five Visualization Panels

Cluster Panel: We designed our first panel representing the articles as a clustered bubble chart, with all of the similar articles grouped together into clusters, each with a distinct hue signifying the most commonly used propaganda technique. The bubbles were sized according to their propaganda content which we received from PRTA API response and binned according to the largest summation of weights received. We added a hover tooltip over the bubbles to show the publication, the date and a pie chart with the various propaganda techniques and its distribution in the article. When the user clicks on a bubble the article text appears in the Article Panel and the Word Cloud Panel shows its corresponding word cloud. *Streamgraph Panel*: We designed the streamgraph panel to depict the propaganda techniques used in the shown articles with respect to time. For convenience, the colors in the streamgraph correspond to the colors in the cluster panel. When the user hovers their mouse over one of the streams, a tooltip appears with more information about the propaganda technique's application. *Article Panel*: We designed this panel to appear when user clicks on an article in the bubble chart. This panel shows the entire article text with the entities highlighted in color by Named Entity Recognition types using Spacy library to calculate the entities and then circumventing them into span elements. *Wordcloud Panel*: We designed this panel to show the most frequently used words in a particular article after removing stopwords using NLTK. We also added a hover tooltip to show how many times a word has been used in the article. *Motionchart Panel*: We designed this panel to show how the political leaning of newspapers changed over time. We binned the articles by publication and then sorted according to date and then displayed the difference between the left and right biases returned by the response of the API. The Y axis is scaled from -1 to 1, with +1 indicating the most left bias and -1 indicating the most right bias of articles.

We had to preprocess the Kaggle dataset and convert it into json files containing article titles, publications, dates and propaganda scores, bias scores supplied by the PRTA API. The data was then integrated into the visualization panels after it had been handled. For article retrieval, we used the tf-idf (bag of words algorithm) to convert the content of each article into vectors, and we used the vectorizer to convert the search criteria given by the user into the same vector space. Following that, using cosine similarity as the similarity metric, the top related articles are obtained from the dataset.

III. EXPLANATION OF RESULTS

The backend algorithms and PRTA scores accurately describe the articles. PRTA propaganda scores will be greater if the articles contain a lot of propaganda. The API's bias scores are organized into three categories: center, left, and right, with a total sum of scores equal to 1. If the articles are left prejudiced, as evidenced by the content, the articles are truly biased, and vice versa. We noticed that the model is rather accurate, albeit it does highlight some of its limitations. Only 120 calls to the interface could be made on average per day, with an extra constraint of 30 calls per hour, prompting us to decide to parallelize article scoring.

The visual system we devised had some intriguing outcomes. The clustered bubble chart revealed noteworthy similarities in the articles in the same cluster, such as comparable language and propaganda strategies. On the basis of the same rationale, an article propaganda method may be predicted as a future scope. The streamgraph revealed the quantity of propaganda utilized in articles through time, with some articles using far more propaganda techniques than others during a year, possibly due to a unique event that occurred that year. The motion chart panel illustrates a fascinating trend of media shifting their bias over time. We can observe that some media choose to be left biased in some years and right biased in others. Article bias could be influenced by an external element. The article highlighted with Named Entity Recognition types also reinforce the reliability of our system with the users and they can use it to look into the specific phrases that have contributed to the article’s biasing.

Biased reporting can be easily identified by professionals in the domain as well as the general public due to our system’s easy to use features, which makes this type of technology usable for many fascinating. Our system could be used for the following potential use cases:

Reporting patterns in news organizations' propaganda techniques: There is a long history of media bias in American news reporting. Biased reporting has major societal implications. Analysts and academics can utilize our method to support claims of unbalanced reporting. They can also figure out the reporting organization's true

method. DellaVigna and Kaplan identified evidence in a 2007 study that media bias (especially for right-leaning Fox News) influenced voting in the presidential elections of 1996 and 2000. According to a 2015 study, biased news can cause dissent intolerance, political segregation, and group polarization. Today, several watchdog organizations, such as AllSides and the Center for Media and Public Affairs, investigate media bias and framing. While these groups often include some qualitative evaluation, academic research focuses mostly on data-driven methods for assessing bias in news reporting. [4]

Analyzing political inclination changes of publications with changes in administration: Researchers can use our system to look at how a news organization's political tendencies changed as the administration changed. If the reporting organization is pushing a narrative in favor or against the present government, analysts can utilize a bubble chart along with a named entity recognition panel to back up their argument. Similar articles with similar political agendas can be viewed in the bubble chart and their summaries can be drawn up through our system. The publications with respect to time keep changing their political inclinations as shown by the motion chart. For example, during the Trump election, articles in favour of his candidacy might have changed their inclination at a later stage during his presidency. This could easily be analyzed by looking at the motion chart panel which depicts the shift in biases over time.

Choosing the most successful propaganda strategy for a certain use case: Researchers can examine news organizations' shifts in propaganda strategies over time using the stream graph generated by our system. Researchers can also utilize our system to see which propaganda tactic has been most effective in the past for a specific use case. Consider the election as an example of a specific use case. If we look at the papers about Trump's first speech, we can see how he repeated a few phrases, such as "make America great again," and how he used specific jargon to grab Americans' attention and trust. This could be used as a tactic for future elections as well, and who knows, it might work.

IV. MY CONTRIBUTION

The Cluster Panel and the Motionchart Panel are two UI and visualization panels on which I worked. First I made a list of keys to address the different Propaganda techniques like ‘loaded language’, ‘namecalling’ and so on. Then I sorted them and initialized a colorScale based on the same color codes as the Streamgraph. The data received from the backend included the propaganda scores which I used to cluster the respective articles. I used d3.forceSimulation wherein a cluster force pulls each cluster of nodes toward its weighted centroid, while a collide force keeps nodes from overlapping. The latter is a custom d3.forceCollide implementation that employs different distances to separate nodes belonging to the same group against nodes belonging to different groups. [5] To create the nodes in the svg, I added circles and added a tooltip for future use. I calculated the centroids of the nodes based on the x and y values of articles along with their respective radii and generated the functions d3.forceCluster() and d3.forceCollide() to attract nodes closest in distance to their centroids and to prevent overlap of circular nodes respectively. I was also in charge of the Motionchart Panel in addition to the Cluster Panel. I added the x and y axes to the svg for the Motionchart, plotted circles for publications, and displayed the year corresponding to it. The circles are redrawn and the transitions are based on bias values interpolated from the data for the fractional years as well. Each time the last year in set is reached, the animation is repeated.

V. LESSONS LEARNED

I learned how to use the fascinating PRTA API to determine the score of propaganda techniques employed in articles from a variety of newspaper sources. Initially, I planned to cluster the articles using KMeans, but this would introduce an unnecessary problem of overlapping, which is easily avoided by utilizing D3's Force Layout to create a clustered bubble chart. Using this, the animation was easier to implement and more adaptable. Based on laws of physics, the force simulation uses force functions, which adapt the location and velocity of objects to produce effects such as attraction, repulsion, and collision detection. I also learned how to use D3.js to construct an animated motion chart and how to integrate data from APIs and generated json files into the visualisations in JavaScript. I learnt about the concept of Promise in JavaScript and how the data can be handled and pre-processed before presenting it to the user. The whole experience was great, although the learning curve was steep due to my lack of experience in front-end, but well worth it because it taught me a lot.

VI. TEAM MEMBERS

Ananya Pal

Aniket Devle

Chirag Vartak

Kevin Shah

Saikat Datta

Siddhant Srivastava

VII. REFERENCES

[1] Martino, Giovanni & Shaar, Shaden & Zhang, Yifan & Yu, Seunghak & Barrón-Cedeño, Alberto & Nakov, Preslav, "Prta: A System to Support the Analysis of Propaganda Techniques in the News," 2020.

[2] "Tanbih API," [Online]. Available: https://app.swaggerhub.com/apis-docs/yifan2019/Tanbih/0.8.0.

[3] "Kaggle All The News Dataset," [Online]. Available: https://www.kaggle.com/snapcrack/all-the-news.

[4] Aditi Mishra, Shashank Ginjpalli, Chris Bryan, "News Kaleidoscope: Visual Investigation of Coverage Diversity in News Event Reporting," 2021.

[5] "Clustered Bubbles," 2019. [Online]. Available: https://observablehq.com/@d3/clustered-bubbles.