**Beth Schueler Research Context (John Synthesis)**

In recent years, education has become a particularly contentious battleground in politics, especially in the state of Virginia. The major point of contention has been the handling of COVID-19. School districts were faced with tough decisions, ranging from keeping schools entirely virtual to reopening schools and completely foregoing masking. While the handling of the COVID-19 pandemic has been one area of contention, another set of escalating concerns have been “Identity Politics”— *Critical Race Theory* (CRT) and *LGBTQ+* issues. In particular, there is concern among some parents that schools are teaching students that “whites are racist and blacks are victim” or that schools are enabling “pedophilia homosexual leanings and pornography books” in libraries. The governor candidate at the time, Glenn Youngkin, actively made those contentious topics a focal point in his campaign during the time leading up to the election.

While we [Beth] know that politics affects policy, it is less clear how policy affects politics. To *causally* answer the question of how policy affects politics is challenging: policies are non-randomly chosen and implemented, so their influence on politics is often endogenous. However, it may be possible to understand the role of policies on politics if there is variation in the exposure of policies to units of analysis that are otherwise identical in observable and unobservable characteristics. Spatial Regression Discontinuity Design is one approach— individuals that live on one side of a cut-off (e.g., county) may not be systematically different than individuals that live on the other side of a cut-off. Although education (school) ‘quality’ often motivates the choice to live in a county, the explicit education policies of a particular school district are less likely to be salient in the decision of where to live.

Data on school districts’ COVID-19 policies and voting at the *precinct­*-level for the governor election were collected. Precincts that border each other but are in different counties are arguably identical to each other *except* for their exposure to observable school district policies. By exploiting the variation in education policy for otherwise identical precincts, a *causal* estimate of how policies affect politics is attainable.

**John’s Role**

While COVID-19 policies have been systematically collected as part of a larger research effort for the Virginia Department of Education, policies related to *Critical Race Theory* and *LGBTQ+* issues have not been. These policies often do not actually exist at the school district level in Virginia, so there is little variation in the policies. However, these issues still appear salient with many attributing Youngkin’s victory to his focus on contentious “Identity Politics”.

To better understand the salience of “Identity Politics” in schools, we leverage data from the public comments section of school board meetings. A team of undergraduate research assistants systematically collected videos of the two school board meetings leading up to the November governor election (11/2/21) for each school district in Virginia. The videos were transcribed using otter.ai to transform the data into text. When video data were unavailable, transcripts of school board meetings along with the public comment summaries in meeting minutes were used. 114 out of the 132 school districts had at least one usable artifact—87 had video(s), 2 had transcripts, and 25 had public comment summaries.

We theorize that the salience of “Identity Politics” in schools may manifest directly in the public comments that community members choose to discuss in the school board meetings. In essence, we hope to leverage public comment transcripts to derive a *latent* construct of the salience of “Identity Politics” in a particular school district. The underlying salience of “Identity Politics” construct may moderate the relationship between COVID-19 policies and voting, so it is imperative that it is included in the modeling strategy or otherwise bias the causal estimates.

*Analytic Plan* [TOOLS 🡪 Establish my Corpus 🡪 Build into the data model we talked about 🡪 Do the different models we covered in class 🡪 Exploratory phase 🡪 Make observations and notice patterns that may eventually be discussed in the paper. Show that you have gotten some information from those models. Not really changing the representation of text. Same kind of stuff from different angles. Recurring themes that occur in the model and that’s what I talk about in the final product. Relationship between Method and Research Question inverted 🡪 focus on method.

* Data Structure
  + Naming Scheme
    - 0\_RAW\_DATA
    - School District [e.g., Accomack County Public Schools (001)]
    - YYYY.MM.DD\_Transcript\_School\_District (e.g., 2021.09.21\_Transcript\_Accomack\_County)
  + Transcript Structure
    - Unknown Speaker X:XX, XX:XX, X:XX:XX, where X’s are the time.
    - Text of the speaker follow the unknown speaker time with a line break.
    - \*Notes
      * All transcripts have been *STANDARDIZED* to match otter.ai
      * While otter.ai is capable of identifying *unique* speakers, it is not great at identifying whether the same set of individuals are engaging in a back-and-forth.
        + For example, Person 1 might ask “Can you hear me?” and Person 2 will say “Yes, we can hear you”.
        + If Person 1 proceeds with speaking afterwards, they will be treated as a distinctive “Unknown Speaker” even though they are the same person. As a consequence, otter.ai overcounts the number of unique speakers.
      * In contrast, the speaker summaries vary quite a bit.
        + Some school districts summarize each individual speaker, which allows seamless standardization with the format of otter.ai.
        + Other school districts may indicate something to the effect of “12 speakers talked about CRT”.
        + Speaker summaries, on average, undercount the number of unique speakers.
      * Some transcripts and summaries did not have proper times associated with them, so I approximated the times.
        + I start the first speaker at 0:00
        + I place the text of the public comment into wordcounter.net – the website has a feature called “Speaking Time” that estimates the amount of time a text will take to be read aloud at a rate of 180 words per minute.
        + The “Speaking Time” of the text is rounded up to a :15 second interval (e.g., First speaker has an estimated “Speaking Time” of 1 minute 33 seconds, so I round up to 1 minute 45 seconds).
        + To also account for transition time, speakers are given an additional :15 seconds (i.e., the total “Speaking Time” for the first speaker becomes 2 minutes).
        + For obvious back-and-forth with small comments in manually standardized school district transcripts (Caroline County Public Schools) I use a :05 second change instead.
* Data Models
  + Ordered Hierarchy of Content Objects
    - School District
      * Time 1/2 Transcript
        + Unknown Speaker X:XX

Sentences

* + - [school\_district, transcript\_time, unknown\_speaker, sentence]
* Exploratory Text Analysis
  + Describing the Data [Zipf’s Frequencies + TF\_IDF/DF\_IDF]
    - Occurrence and Importance of words
      * Zipf’s: Global Term Frequencies
      * BOW & TF\_IDF: Co-Occurrence. Weight terms by multiplying their raw frequencies by the inverse of their frequency in the document corpus as a whole. These shed light on **corpus-level** term significance when aggregated (mean) at the Vocab level.
      * \*For describing the data, these terms will help us identify what is “salient” for the overarching dataset. Since I have a choice of what BOW to employ, I think that the school district level is the group that I am interested in.
    - What do our important words mean? What are they connected to?
      * **Word vectors** generated by a class of **unsupervised algorithms** for learning **the meanings of words** from their contexts.
      * Can use this for analogies.
  + School District Similarities / Clustering
    - GROUPING OF SCHOOL DISTRICTS [Hierarchical Agglomerative Clustering] 🡪 “Genre”
      * Cosine Similarity is often used because it’s already normalized for length [this may be valuable since we know our transcripts vary greatly in length from one speaker to over an hour of speakers].
      * “Complete Linkage” focuses on global qualities of clusters. It produces more balanced clusters. It is less susceptible to noise. Often breaks very large clusters. 🡪 preferable for language modeling. Ward Linkage also might be appropriate and typically used for hierarchical clustering.
    - UNDERSTANDING WHY SCHOOL DISTRICTS ARE GROUPED 🡪 “Style” [?How useful is “Style” for my purposes?]
      * Principal Component Analysis
        + PCA: “Sheds light on why the texts cluster as they do” – Exposes **exact lexical differences** that constitute difference.
        + \*May be slightly useful to identify how districts fall along principal components… but I think it’s unclear to me as to how many components to display / the school districts might be grouped along many different dimensions.
      * Latent Dirichlet Allocation
        + Probabilistic model that tries to estimate the probability distribution for topics in documents and words in topics
        + Helps surface latent semantic structures across collections 🡪 collections of documents as bearers of cultural information
        + Use Collapsed Gibbs Sampling
        + Heteroglossia: Multiple perspectives within documents 🡪 seems definitely relevant considering we have many different speakers
        + Alpha parameters as a table of contents 🡪 “Topic **alpha** is closely related to plain old topic **frequency** – the number of times a topic appears in a document above a certain threshold (something like 5%)”
        + “Topic Entropy” – Degree of topic randomness in a document. Document about one thing has low entropy; document about multiple things have high entropy [I can imagine this being the basis for sentiment analysis stuff?]
        + “Topic Trend” – average weight of topics in documents of a given date range 🡪 In theory this could be done with the speaker order, but seems low priority.
        + “Topic Similarity” – Calculate the topic’s distance in vector space 🡪 can we reduce down the topics in this way? Or does it make sense to tweak parameters such that we reduce the number of topics.
        + “Topic Contiguity” – Measure of topic’s co-occurrence across a corpus 🡪 If one topic occurs, then how likely is it for another topic to appear? [Again, maybe this is a way to identify a potential reduced down set of topics? 🡪 But I could also see a particular POV emerge with Religiosity + COVID-19 + CRT + LGBTQ+]
        + “Topic Networks” – another way of visualizing topics. Less interesting for my purposes unless we can derive factor loadings.
      * Do these school districts ALSO have similar levels of sentiments on these particular topics?
        + \*Do this at the individual speaker level. Topics + Speakers.
        + Q: Does it make sense to simply do topic modeling on the speaker level and aggregate it to the district? If I do it at the district-level, I’m concerned about conflating the district’s overall sentiment on unrelated topics.
        + Feels like the bulk of my figures will be from here.
      * Word Shift Graphs
        + “Which words contribute to a difference” between two texts and how they contribute 🡪 this might be a useful tool for understanding differences between districts, especially with respect to topics. IS THIS REDUNDANT?
* Beth Analysis
  + **Question: Do we have to use everything we learned? It appears that word2vec has the least overlap with the goal that I am trying to accomplish**.
  + Set-up for “Perlocution” (Impact Analysis) using the “Utterance”
    - The “Locution” (Topic Modeling)
      * Since the analysis is at the precinct-level (school district is closest we can approximate this), we need to identify the global topics being discussed in the entire school district.
      * First, we want to uncover the broader thematic topics that are being discussed in the broader corpus.
      * We would then want to identify the “Topic Entropy” 🡪 this will allow us to uncover the degree to which a particular document (**Q1: Can I aggregate the separate transcripts into one document for this?**) has a mixture of topics.
        + Low entropy document contains few topics.
        + High entropy document contains more topics.
      * We also might want to look at topic contiguity – despite featuring different words, it seems very possible to me that CRT and LGBTQ+ ideas are closely related in the vector space.
      * Maybe visualize in Topic Network?
    - The “Illocution” (Sentiment Analysis)
      * The reaction toward the “Identity Politics” topic matters. If every single commentor within a school district praised CRT/LGBTQ+ policies, then we might expect this to lead to a decreased vote share for Youngkin.
      * Sentiment analysis therefore serves as a useful tool to understand the degree to which a person positively or negatively reacts to certain topics 🡪 The aggregated results therefore would inform the emotion associated with the topic at the district-level.
      * Frequentiment? 🡪 NO
      * **Question: While we know that sentiment analysis can be applied to the document – whatever level that may be – is it possible to ascertain the sentiment towards particular topics? Logically, I think it should be possible.**
        + **You might first have a set of speakers. These speakers either talk about the topic or they don’t.**
        + **If the speaker doesn’t talk about a topic, maybe we assign them a neutral (0) toward a particular topic?**
        + **If the speaker does talk about a topic, then we assign them a sentiment score overall?**
    - At what level do we conduct this analysis?
      * District-Level
        + If we conduct this set of analyses at the district-level, then we simplify many portions. In particular, we should be able to ascertain to what degree does a particular school district talk about “Identity Politics”.
        + In addition, we can do an overall “district” sentiment analysis.
        + HOWEVER, this feels potentially problematic. The overall district discussions about Identity Politics feels fine and working as intended, but the sentiment portion feels unclear. Is it the topic’s that an individual is reacting to? Or is it some other topic that is being discussed in the district.
        + It feels unclear to me how to associate a district-level topic analysis and the underlying sentiment for those topics unless an explicit tool exists to connect topic and sentiment of the topic
      * Unknown Speaker Level & Aggregate?
        + If we conduct this set of analyses at the speaker-level, then we get more precise derivations of sentiment.
        + In particular, I can imagine the ability to identify a topic as being very clear. The topic entropy then unveils the mixture of which a particular topic belongs to an individual speaker. I could imagine establishing a threshold of X that states this speaker is focused on, say, “Identity Politics”.

<https://blog.insightdatascience.com/topic-modeling-and-sentiment-analysis-to-pinpoint-the-perfect-doctor-6a8fdd4a3904>

[Someone did the threshold thing here]

* + - * + The sentiment of the speakers that focus on “Identity Politics” can then be computed.
        + The lingering question would be how to aggregate this information? Would it simply be the average? I can mathematically imagine something to the following effect:

Identify the degree to which topic is discussed for an individual speaker. If \*above\* threshold, use sentiment score. If \*below\* threshold, use 0. [I imagine that there is a distribution for whether the speaker contains a particular topic. Some threshold might make sense 🡪 it may also be the case that I will need to use a ‘latent’ topic that connects multiple “Identity Politics”]

Rather than focus on limiting the sample to those that discuss a particular topic and their respective sentiment, we still want to understand what’s happening in the overall school district. I think this can be accomplished by the approach of only using the sentiment if a speaker sufficiently discusses a topic.

If we average the topic sentiment scores across an entire district, then we obtain a sense of what the average sentiment is on a particular topic for that particular district. Implicitly if we do not do a weighted average, then we are equally weighting each speaker.

\*Is there value in weighing the number of speakers? Maybe not since district size could easily be integrated into the regression model.

* + - * This document is about this topic 🡪 every topic is theoretically representative in a document. Identify a cut-off, then subselect which documents most represent a given topic. Topics strongly correlated with documents. Can run sentiment analyses on the topics.
        + First pass: this topic seems to be strongly correlated with this sentiment through these documents – based on a heatmap.
        + Can subselect further – group documents by person [or school district]. Set of documents strongly associated with a topic. Which author and do sentiment analysis at that level.
        + Subset of documents that pertain to a particular topic. Those documents have meta-data that says which person produced the document and which district they come from. Assign a sentiment score to each particular document.
        + Quantitative Psych based on NLP – Ask Hudson Golino.
        + Methodology – Correlating Gender w/ Topics.

*Deliverables*

* Data Files
  + A collection of data files, each in CSV format, containing the F2 through F5 data you extracted from the corpus. These files should include, at a minimum, the following core tables:
    - LIB.csv - Metadata for the source files.
    - CORPUS.csv- This is a tokens table annotated with statistical and linguistic features, such as TFIDF. It should include an index that represents the OHCO of the documents in your corpus.
    - VOCAB.csv - Annotated with statistical and linguistic features, such as DFIDF.
  + Features in Appropriate Core Tables or as Separate Tables [Note: all tables should have an appropriate index and, where appropriate, an OHCO index.
    - Principal Components (PCA)
      * Table of documents and components
      * Table of components and word counts (i.e., the ‘loadings’), either added to the VOCAB table or as a separate table with a shared index with the VOCAB table.
    - Topic Models (LDA)
      * Table of document and topic concentrations
      * Table of topics and term counts, either added to the VOCAB table or as a separate table with a shared index with the VOCAB table.
    - Word Embeddings (word2vec)
      * Terms and embeddings, either added to the VOCAB table or as a separate table with a shared index with the VOCAB table.
    - Sentiment Analysis
      * Sentiment and emotion values as features in VOCAB or as a separate table with a shared index with the VOCAB table.
      * Sentiment polarity and emotions for each document.
* Code Files
  + The Jupyter notebooks used to perform all operations that produced the data in your tables
  + Any Jupyter notebooks used to explore and visualize the data in preparation for your final report.
  + Any Python files you wrote to support your work
  + Any other assets – e.g., images, stylesheets, JavaScript libraries, etc. required by your notebooks
* Report Document
  + A Jupyter notebook called FINAL\_REPORT.ipynb – describing your work and interpreting its results along with links to all the files listed above. This report should be written using Markdown text cells and embedded graphics from your other notebooks to illustrate points. Do not reference images that are not listed in the notebook. You may use images to show images in the notebook if you don’t want to include the code there. Include citations for any references made in the notebook.
    - **Introduction** – Describe the nature of your corpus and the question(s) you’ve asked of the data.
    - **Source Data** – Provide a description of all relevant source files and describe the following features for each source file:
      * *Provenance*: Where did they come from? Describe the website or other source and provide relevant URLs
      * *Location*: Provide a link to the source files in UVA Box.
      * *Description*: What is the general subject matter of the corpus? How many observations are there? What is the average document length?
      * *Format*: A description of both the file formats of the source files (e.g., plaintext, XML) and the internal structure where applicable.
    - **Data Model** – Describe the analytical tables you generated in the process of tokenization, annotation, and analysis of your corpus. You provide a list of tables with field names and their definition, along with URLs to each associated CSV file.
    - **Exploration** – Describe each of your explorations, such as PCA and topic models. For each, include the relevant parameters and hyperparameters used to generate each model and visualization. For your visualizations, you should use at least three (but likely more) of the following visualization types:
      * Hierarchical Cluster Diagrams
      * Heatmaps showing correlations
      * Scatter Plots
      * KDE Plots
      * Dispersion Plots
      * T-SNE Plots
    - **Interpretation** – Provide your interpretation of the results of exploration, and any conclusion.
    - Regarding number of pages, a rule of thumb would be a six page exported PDF. The question of length is secondary to the requirement that you answer complete all the sections.
* Format and Style
  + Any non-data files you produce, such as a Jupyter notebook or a Python program, should contain a header stating your name and email address, the name of this class (DS 5001), and the date. It should look something like this:
  + John Wang ([are2ag@virginia.edu](mailto:are2ag@virginia.edu)) DS 5001 Spring 2023
  + Jupyter notebooks should be properly outlined with headers and explanatory text where necessary to follow what is happening.