Replication Project Step 1: Paper Selection

Giovanni Maya

2025-05-02

Paper: Introduction and Argument

The paper selected for my project is titled **Structural Topic Models for Open-Ended Survey Responses** by Roberts et al 2013. This piece has been published on multiple notable journals such as the Midwest Political Science Association, which is where I sourced it from. A quick synopsis of the paper's interest will be useful for our analysis of the data and methods shown in future sections.

The collection and analysis of open-ended data in political science is infrequently used, and when it is employed, analyses of the sort rely specifically on human coding to determine trends (thus, no automation). They mention that the major advantage to analyzing open-ended data is that it provides a significant and direct view into a respondent's own thinking rather an interpretative understanding of their thinking. In other words, it allows for the respondent to express a subjective view that would otherwise be constrained or guided if it was a close-ended question. However, they posit two main concerns that arise as a result of this analysis. For one, there is the challenge of a tendency for subjects to articulate a response instead of their underlying attitudes. Secondly, they have been considered more difficult to analyze than close-ended questions, often requiring various stakeholders to come to a formidable and interpretable output. Nonetheless, the authors note that though these concerns are valid, they should not be overtly substantial in the inhibition of *true insights*. As a result, the core of their **thesis** is the structural topic model (STM) that offers a powerful and flexible unsupervised framework for analyzing open-ended survey responses and systematically relating them to respondent-level covariates, mitigating the need for hand-coding and labeled data.

Datasets

For this project, three datasets were included with the paper, of which I will cover one by one. The main and primary dataset comes from the 2008 Anes Time-Series Study which includes open-ended responses to the question of the most prominent political problem. This dataset includes demographic metadata such as age, education, and of course, the respondents' description of what they believe to be the most pressing problem facing the country. This also included respondents' responses about what aspects of a presidential candidate they liked or disliked. The dataset contains over 2,000 open-ended survey responses paired with 64 columns of detailed political and demographic metadata. A sample is shown in Table 1.

The second dataset was sourced from a political psychology experiment brought forth from Gadarian and Albertson where respondents' responses were gathered on how negatively valenced emotions influence political attitudes towards immigration. The respondents were randomly assigned to one of two conditions. The first condition was the fear-primed condition where respondents were instructed to write about immigration in a way that made them feel worried or afraid. The second condition was a neutral/reflective condition where they were instructed to write about immigration in a more deliberate, cognitive tone. This dataset contains 352 observations and 16 columns describing demographic and experiment encodings. A sample is shown in Table 2.

Age	PID Summary	Issue 1 (mippol1)	Issue 2 (mippol2)
35	4	the economy	immigration
39	5	how our government handles our economic crisis	gay marriage
50	3	the spending budget	education
72	6	economy no	terrorism
66	4	to do what your profess...	its a problem that we have political parties
56	4	morals	morals

Table 1: Sample of ANES Respondent Responses

Case ID	Treatment	Fear (ra1)	Republican ID	Open-Ended Response
287	1. worried	1	1.000	problems caused by the influx of illegal immigrants...
145	1. worried	1	1.000	i'm afraid of who might be getting into this country...
159	0. think	0	0.333	they should enter the same way my grandparents did...
421	0. think	1	0.500	legally entering the usa meeting the requirements is the law...
224	1. worried	2	0.667	terror, bombings, killing us, robbing america

Table 2: Gadarian Sample: Metadata and Immigration Responses

The last dataset was sourced from the RAND Experiment that was designed to study how people reason about tradeoffs in group-based financial scenarios. The main column of interest is the `strategy.description` column that outlines open-ended natural language responses explaining why a respondent chose to contribute (or not to contribute) to a shared pot in a public goods game. The answers reflect nuance and considerable variance between respondents. This dataset contains 150 responses and 42 columns that share similar features as the previous two datasets. A sample is shown in Table 3.

Age	Primed	Contribution	Normalized	Risk	Strategy Description
22	1	0	0.0	4	to maximize my personal earnings
20	1	0	0.0	5	i could not trust that the other group members...
18	1	400	1.0	4	benefits the group the most
21	1	0	0.0	4	i would gain more by not adding money...
18	1	400	1.0	6	i am banking on the fact that people will be less selfish...

Table 3: RAND Sample: Contributions and Strategy Explanations

Unsupervised Methodology

In this paper, unsupervised learning is used as an automated way to extract latent themes or topics from the open-ended responses across different political science experiments. The algorithm they propose is a structural topic model which is an advanced form of the Latent Dirichlet Allocation that incorporates document-level metadata into the topic discovery process and informs more nuanced topics. This method extends LDA by essentially allowing topic proportions to vary with covariates. This is done using the `stm` package in R that both infers the number and content of best topics from the text data.