Condensing Political Opinions

Alternative approaches to word clouds

Alec Jeffery
Student author
112 Mercer St
University of Colorado
XXX-XXX-XXXX
Jefferya17@gmail.com

ABSTRACT

In this paper, discussion of alternative approaches for aggregating and visualizing corpuses of polling data and political opinions. The intent is to demonstrate a novel approach for dimensionally scaling qualitative data to infer a population's opinion distribution.

Large Language Models [LLMs henceforth] can be utilized to compress input text to a shortened version. These compressions can be tokenized and evaluated with unsupervised learning, specifically cluster analysis. Cosine similarity will be utilized to estimate relative distance(s) to other tokenized text and a distribution can be formed by measuring individual encodings relative to the population mean.

Insights about the population and individual datapoints can be visualized in a manner that expands depth beyond typical analysis such as word clouds or frequency analysis.

KEYWORDS

Polling, LLM, NLP, dimensional scaling, political, opinion, word cloud, sentiment analysis, token embedding, clustering, kmeans, DBscan.

1. Introduction

Opinion polls and survey data generate dense corpuses of text data that is typically funneled through trained personnel to process the data. The objective of processing the data is to gather insight to where a population sits relative to a topic or subject. Sentiment analysis has been deployed on large datasets [1] to gauge temperature and word clouds [2] are commonly used to visualize word frequency.

Measuring dispersion of opinions within a poll or survey offers a challenge that is investigated in this paper. A single entry of text response within a larger poll or survey can be compressed with an LLM to a shortened version of the text. The investigation of this project will deal with how these token embeddings relate to one another and the larger population of responses.

Expanding upon traditional analysis of polling data offers novel use cases for policy makers and candidates. For better or worse, candidates typically adhere to a strategy of targeting one end of the opinion spectrum. Consequently, issues affecting the tails of the political spectrum get the most attention in public discourse. This paper presents an approach for quantifying the center of an opinion distribution. The significance of this approach is that engagement with a population can be shifted to more moderate stances and facilitate a cooling of political vitriol.

2. Related Work

In the realm of compressing and visualizing an opinion dataset there are notable approaches. DW NOMINATE [3] is the primary tool for scaling voting data into a dimensional reduction. A congressional member's votes are compared against their respective party and a distribution of how various members differ from the central tendency of the party as a whole [**below**]. Lower-level representations of such data allow observes to quantify the behavior between groups and individuals.

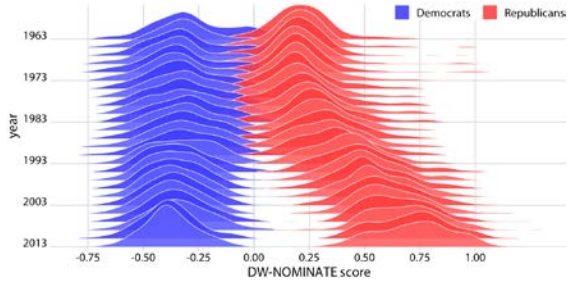


Figure 1: Visualization of Congressional voting.

In terms of representing rich text responses to survey and polling data a common practice is to utilize word clouds. This approach has limitations because it strictly considers the frequency of words but does not reflect sentiment associated with the words. For example, the name "Trump" may occur frequently in a discourse but opposing viewpoints of "I like Trump" as opposed to "I loath Trump" will not be differentiated in a word cloud or frequency analysis.

More recently, LLMs have entered the sentiment analysis space and offered new approaches. Instead of machine learning approaches for sentiment analysis, DNNs can be used to classify sentiment in a body of text. Similarly, Google llc, developer of BERT [4], has shown promise in DNN seq2seq compressions via

transformer architecture. Dimensional reductions that leverage seq2seq offer a more parsimonious approach to sentiment analysis.

3. Proposed Work

This paper explores approaches for dealing with text data as applied to polling and survey data. This work will focus on a qualitative polling dataset from the 2016 General Election of religious participants was performed. The individuals in the poll gave free text responses to specific questions as it relates to their opinion. The data is in .csv format and is imported to python via pandas. The data has been cleaned by tagging and removing partial responses from the dataset. Since an LLM is being used to perform a compression, common 'stop words' will not be removed from the statements. Upon data cleaning & preprocessing, there are 2195 rows of data.

Because there are multiple topics as well as multiple individuals, this dataset will be insightful as to whether the responses can be aggregated along a spectrum of possible opinions. Moreover, developing a distribution will be highly insightful to demonstrate a central tendency of the responses. Lastly, a visualization that shows where a response resides relative to the mean group response will help demonstrate extremes in opinions.

This work will be predominantly performed within a Python Jupiter notebook. Due to limitations of memory on a local machine, OpenAI API will be leveraged to compress opinions in the poll data. The technique will specify a maximum number of characters for the compressions and then leverage an open source LLM to perform the token embedding of the compressed data. Alternatively, cosine similarity can be used to compare embeddings with one another.

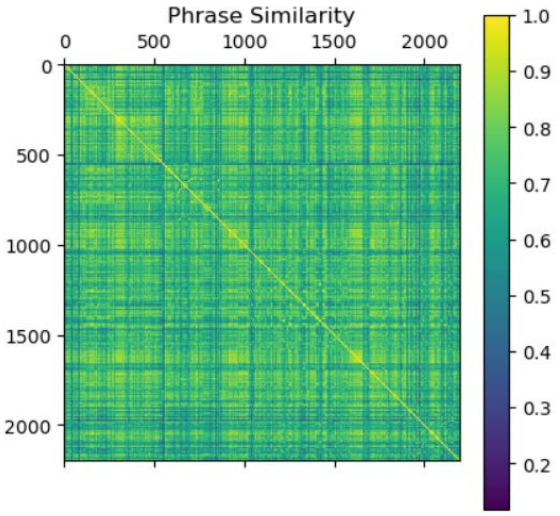
Exploratory data analysis can be performed at this stage. Unsupervised learning with clusters will be performed on the distance matrix of datapoints. K-means, agglomerative and DBscan clustering approaches were explored. K-means is the preferred approach for this paper on the basis that the method generates a cluster centroid which can be used for further analysis.

Cluster centroid can be measured so that the distance of individual datapoints to the centroid will be the basis for gauging the distance from the median opinion of an opinion. A histogram of the distribution will be an informative data visualization. Moreover, sampling of datapoints with low distance to centroids will act as a typical or medial opinion within the corpus.

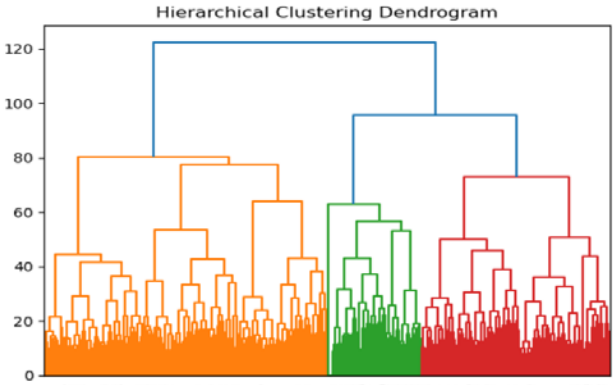
with an established distribution, population statistics can be harnessed. One standard deviation from the population mean will represent a sizable portion of the population. Therefore, sampling from the distribution at +/- 1σ of the population mean will confer insight as to how the middle of the population regards a polling subject. LLMs can be used to synthesize these two bookends of opinions and formulate a composite position that satisfies both ends of the sampling.

4. Evaluation

Evaluating similarity between datapoints within an unlabeled dataset requires qualitative & quantitative approaches. Following data preprocessing and token embedding with the BERT LLM, similarity can be visualized with a diagonal similarity matrix that shows each datapoint relative to one another. Rows and columns are identical, the diagonal of the similarity matrix will be equal to 1.0 indicating perfect similarity with itself. For offset rows and columns a similarity score can be calculated with cosine similarity in the PyTorch library. Lighter colors indicate sections of high similarity whereas darker colors indicate relative dissimilarity.

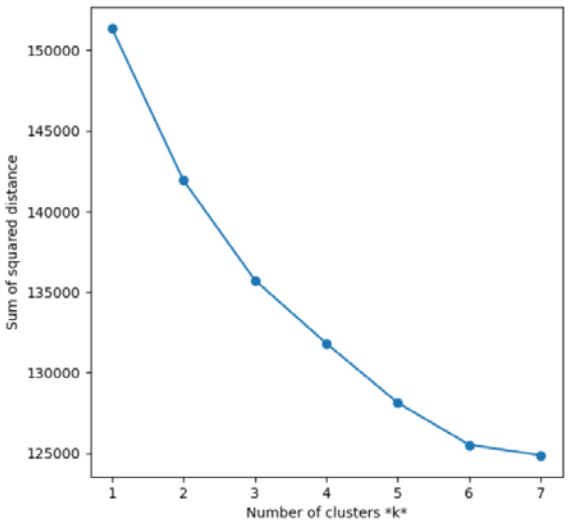


To visualize and evaluate how particular phrases may group with one another, unsupervised cluster analysis was performed. Multiple tools are available for performing unsupervised cluster analysis, for this report SkLearn was utilized. Hierarchical clustering was used to establish boundaries between various topics within the dataset. These boundaries are visibly apparent when presented as a dendrogram.

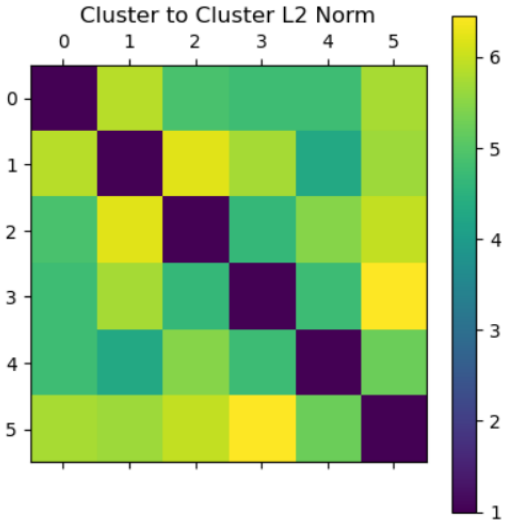


Based on distance thresholds for clusters there are intuitively two to six apparent clusters. Additional approaches must be taken to choose an appropriate number of clusters for subsequent analysis.

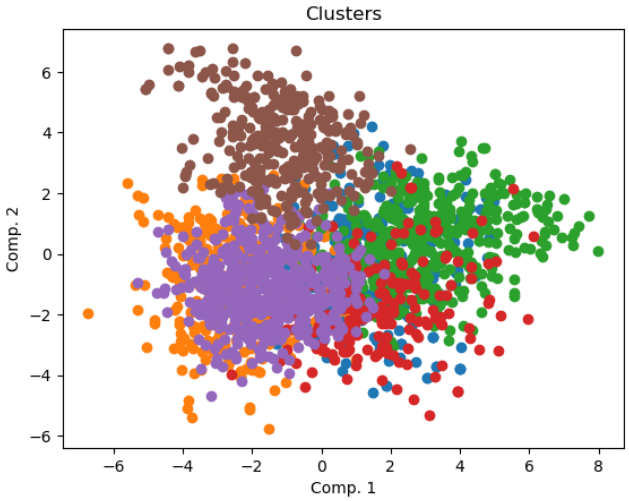
Using K-means clustering to separate the dataset into groups can efficiently be addressed with an elbow diagram. Euclidean distance between cluster centroids was calculated for a given number of clusters. When the total sum of squares distance for all points within a cluster begins to show diminishing reduction for a given number of clusters, an intuitive number of clusters can be selected. In this dataset, the elbow diagram indicates 6 clusters as an acceptable level [Below].



Furthermore, the k-means clusters can be shown with a diagonal matrix in a similar manner to who individual datapoints were previously shown. The L2 norm between clusters is plotted for the six clusters.



Visualizing clusters relative to one another has proven to be challenging. Because the BERT model has 768 token embeddings, each phrase essentially has 768 dimensions. As mentioned before, similarity can be effectively quantified with cosine similarity. Principal component analysis was performed on the embeddings; however, the first 2 principal components describe only 15% of the total dataset variance. Although this is a small portion of the total variance in the dataset, a visualization in lower dimensions is still possible. Below is a visualization of the first two principal components with clusters individually codes in separate colors.

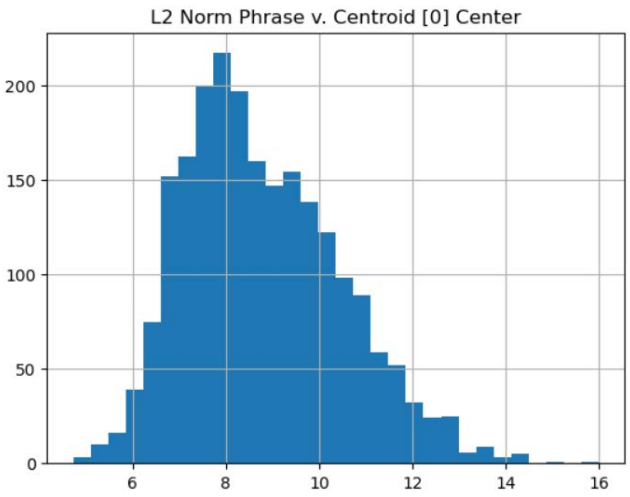


The dataset is sufficiently sized such that server warehousing is not essential to the task. The raw data has been imported with pandas and stored as a dataframe with columns indicating the original phrase, the compression and cluster assignment. Comparing inter-cluster phrases with one another does qualitatively show similarity.

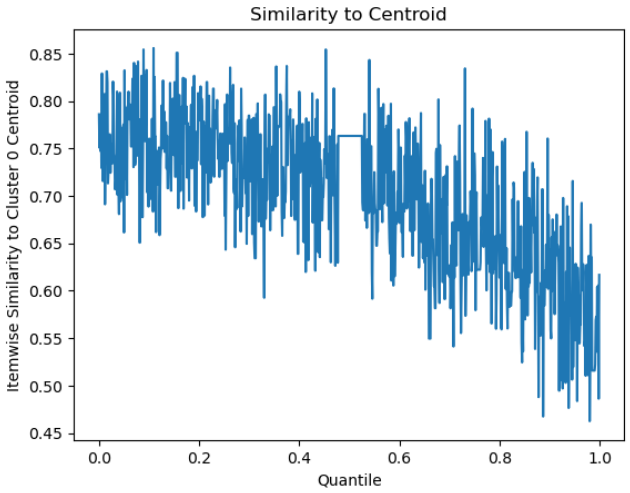
Phrase 1:
"…We have a solid system (government) in place we do not need to falter and weaken and become corrupt as many other countries."

Phrase 2:
"I think it is very important for a President to be a moral leader…."

The distribution of datapoint distance from centroid can be visualized with a histogram. Comparing phrases closest to a centroid versus other phrases closest to another centroid will be informative.



Distance to centroid is useful in terms of quantifying sympathetic viewpoints respective of the specific topic. Quantiles of the distribution can be used to empirically sample original and compressed text. Using two phrases from the original dataset that are each at different quantiles of the distribution provides a bookend of opinions. These two phrases were used to prompt an LLM to generate a new statement that reconciles the two statements. This new statement encapsulates the two statements as well as all statements within the quantiles. The new statement was token embedded and cosine similarity was used to evaluate cosine similarity.



The right tail of the distribution increasingly becomes less similar to the LLM generated statement; whereas the left tail of the distribution tends to be more similar to the new statement. While all points are contained within the same cluster, it becomes more challenging to formulate a composite statement for upper quantiles of the distribution.

5. Discussion

This methodology has been shown to be useful in obtaining the voice of a population's center regarding a specific topic. Topics do not need to be predetermined and are thus unsupervised in this regard. In addition to finding a central voice of a topic, an inter quartile range can be established and LLMs can be used to synthesize a generalized statement that represents the range.

It will be informative to perform this analysis on a labeled dataset and confirm whether unsupervised clustering approaches are adequate. Scaling the approach beyond this dataset will provide insights into the robustness of the methodology. Given this poll's limited subject matter, a broader dataset will demonstrate population trends that are less narrow in scope and interest. Performing a poll with a single broad question and analyzing the data will be a litmus test for the data mining's application for generalizing corpuses.

6. Timeline

This data mining approach took six weeks to complete. A breakdown of the development cycle is detailed below.

Two weeks were committed to data collection and preprocessing. Removal of deviant or otherwise unhelpful data (i.e. incomplete responses or ineffective data importation) will be taken. Initial clusters will be established, and parameter tuning were performed.

Two additional weeks were used to build and refine distance measurements and distribution creation. Data visualizations will be explored and proposed according to feasibility.

The remaining two weeks were to compile a report and prepare presentation materials.

Regarding future work, a standalone model that takes raw text and generates a report will be the next working step. Integrating interactive plots via Altair or other visualization tools will simplify the task and allow for dissemination to a broader audience. The two of these future steps will take a considerable amount of time.

7. Conclusion

This report has detailed a data mining approach that uncovers trends in plain text polling data. This methodology exceeds the content depth of traditional methods and provides insight into the voice of the population.

Compressing text with an LLM prior to token embedding was shown to be an efficient preprocessing technique and reduced the dimensionality of data being used for clustering. K-means clustering was selected as the preferred approach for detecting and visualizing dataset structure. Maximal cluster elements were determined via elbow plots.

With the dataset reduced to a simplified representation within clusters, element-wise similarity metrics were used to quantify a cluster's content. Moreover, an interquartile range was used to obtain boundaries respective to the centroid. A new phrase can be formulated from such a range to represent a quorum of opinions.

This work has opened the possibility of engaging electorate et masse in a novel way. Given an issue or candidate, an approach for analytically extracting a viewpoint that represents a majority slice of the population can easily be derived. This significantly improves upon the paradigm of seeking out a fervent and loyal base with little regard to most of a group. It is the hope of the author that such a method can be harnessed to alleviate the environment of tribalism and opposition politics.

Further work is needed scale the approach and apply it to different datasets. In particular, the approach may lend insight to business problems involving reliability claims. When a product warranty claim is made, a short questionnaire is taken, and customers provide some text based responses regarding product quality. This data can be used to establish frequent issues with product quality.

8. ACKNOWLEDGMENTS

I would like to acknowledge my children who are patient with me as I pursue my Master of Data Science.

REFERENCES

[1] [Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis] (https://aclanthology.org/D15-1303) (Poria et al., EMNLP 2015)

[2] Schubert, E., Spitz, A., Weiler, M., Geiß, J., & Gertz, M. (2017, August 11). Semantic Word Clouds with Background Corpus Normalization and t-distributed Stochastic Neighbor Embedding.

[3] Poole and Rosenthal, Congress: A Political-Economic History of Roll-Call Voting, Table 3.1 p. 28.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, 2019, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding