# Final Report

**Bhargav Narasimha Shandilya**

# Clustering and Semantic Analysis of Twitter Data
# Piper Gradient Project

## 1. Introduction

Over the past decade, social media has been shown [1][2] to trigger the formation of online echo chambers. People within these echo chambers maintain largely homogenous opinions on key political issues. Radical elements within the group continue to push the bounds of fanaticism and identity politics, moving the entire group further away from the center of the political spectrum. The larger aim of this project was to attempt to reverse-engineer the issues caused by social media and present users with more diverse forms of content. I took up the challenge of performing clustering and semantic analysis on Twitter data, the first piece of the puzzle in addressing the issue.

**The objectives for the project can be summarized as:**

- Identifying the most effective means of pre-processing the data
- Identifying clustering algorithms that would be suitable for text clustering in this context
- Performing the clustering exercise and comparing the quantitative performance metrics for each of these algorithms
- Qualitatively evaluating the clustering results and attempting to find relevant patterns within clusters through sentiment analysis

These objectives would result in a preliminary data pre-processing, clustering, and evaluation pipeline that would feed into the next step of the CONOPS of Gradient - which is analyzing and labeling the clusters.

**The challenges for this approach can be broadly be summarized as follows:**

- The political typology suggested by the Pew Research 2017 survey may not map perfectly onto the ideal number of clusters.
- Clusters may not be inherently representative of something as abstract as a political ideology. The algorithm might be looking at an entirely different set of features for clustering.
- On a more trivial note, the textual data is quite large, resulting in the formation of massive sparse matrices post vectorization. Algorithms like k-means may run into memory issues since they require all the data to be loaded into the RAM concurrently. There are natural workarounds for this problem that are discussed later in this analysis.

# 2. Challenge Fundamentals

## 2.1 Exploring the Data

The data was primarily composed of Twitter user descriptions and 111,271 tweets (after dropping duplicates) made in 2021 with hashtags directly relating to the Infrastructure Bill that was being debated in the US congress. Each row of the dataset culminates in a single tweet related to the US Infrastructure Bill. Supplementary information about the user who made the tweet is available in columns such as 'followers_count', 'friends_count', 'favorites', and so on. However, we are only interested in the user description and tweet columns encoded in UTF-8 format (for instance, u'\xe2\x80\x9d' would represent â). Naturally, the sentiment of the tweets is split between those supporting the infrastructure bill and those opposing it.

## 2.2 Literature Survey

Ryosuke Harakawa et al. [4] put forth several different clustering techniques, including k-means clustering, the Louvain method, and sentiment consensus, that could be used with tweet networks. The network-based clustering approaches that the authors suggest aim to exploit the inherent semantics and sentiment within tweets to draw out relationships. My project was heavily inspired by the algorithm comparison techniques and evaluation methods that they employed.

Since I eventually picked different algorithms centered thematically around the k-means approach, Soni et al. [5] served as a guide for much of the process. The authors of this paper also use term frequency-based feature vectors to train a k-means clustering model, similar to the implementation in my project. Although a density-based clustering method (DBSCAN) was suggested in [6] to account for noisy data, the spectral clustering approach remained largely absent from literature related to tweet clustering. This prompted me to take this up as a possible alternative to k-means clustering. Other algorithms for tweet clustering such as c-means [7][8] and hierarchical clustering [9][10] were suggested by Vicente et al. and Ifrim et al. respectively. Furthermore, the work of Miyamoto et al. [11] continued to be a prime inspiration for cascading multiple clustering methods. Furthermore, [12] suggests an entirely fresh approach to tweet clustering based on retweet identification and agglomerative hierarchical clustering, graphically visualized as a dendrogram.

In [13] we see an extended description of the VADER (Valence Aware Dictionary for sEntiment Reasoning) sentiment analysis. [4] showed that it is possible to apply the VADER NLTK package to clustering tweets based on sentiment. [3] served as the primary source for understanding political typology, and [14] was the source for understanding the 2021 Infrastructure Bill.

## 2.3 Implications of the Survey

Over the semester, I hoped to compare different text clustering techniques to cluster tweets belonging to certain political typologies as described by [3]. The Pew Research Survey conducted in 2017 [3] identified several political categories: **Solid Liberal, Opportunity Democrat, Disaffected Democrat, Devout and Diverse, Bystander, New Era Enterpriser, Market Skeptic Reps, Country First Conservative, and Core Conservative**. The challenge is to perform clustering and evaluate the clusters by matching clustering assignments with the political typology labels suggested by the Pew Research Study or the Piper SME labels

(**fringe-left, progressive, democrat, centrist, libertarian, republican, Trump-republican, fringe-right**). Although labeling is not part of my challenge requirement, it is a necessary step for preliminary evaluation.

Following the literature survey, I found that k-means and Mini Batch k-means would be interesting avenues to explore for text clustering. Both of these techniques are fundamentally simple ways to perform clustering with a TF-IDF matrix representation of data. Since we are refraining from the usage of more complex vectorization techniques such as word2vec and GloVe, the overall strategy incorporates intra-cluster sentiment analysis as an additional clustering technique to verify the effectiveness of the clustering algorithms. Both versions of the k-means algorithm are ineffective when it comes to performing clustering exercises on point clouds such as the one shown below. The figure on the left demonstrates k-means clustering while the figure on the right demonstrates spectral clustering on the same data. Moreover, a key step in the spectral clustering process is just a variation of the k-means cluster assignment step. Thus, we stay with the overall theme of distance metric-based clustering techniques throughout the project. (See figure (f) in the appendix for comparison of Spectral Clustering and K-means Clustering for a data cloud)

# 3. Analysis Report

## 3.1 Defining Evaluation Strategies

We begin with data pre-processing, vectorization using TF-IDF scores [15], and then implement one of the three clustering algorithms. From the clusters obtained, we need to determine if there is any semantic relationship between tweets within a cluster. This semantic relationship must subsequently be related to political typology as defined by [3] or the Piper Gradient SME. To achieve this, we perform sentiment analysis on the tweets within each cluster. Positivity and negativity in sentiment serve as representations of support for the bill vs. disdain for the bill. This is done using the NLTK Vader package. The package is sensitive to the usage of internet slang and emojis, making it ideal in this scenario. We can then plot a word cloud (a list of most frequently occurring terms) and see if the tweets within a cluster conform to our definition of a particular political category. For instance, if we see multiple occurrences of terms such as '#BidenHarris' and 'down with Moscow Mitch' (that's actually a phrase in the dataset), we may conclude that the cluster primarily consists of a particular category of democrats.

Alternatively, we could also use the user descriptions to categorize people into one of 8 classes according to Piper Gradient's labeling scheme. The labels assigned post clustering can then be compared with the labels derived from the user description. Most of the evaluation metrics such as homogeneity, V-measure, and Rand Index will not compare the actual labels. They only measure the extent to which pairs of labels co-occur. We also conduct hypothesis tests using MCMC simulation on the test set to see if the sentiment analysis is actually giving us meaningful political typologies.

## 3.2 Data Pre-Processing

Before the data can be clustered, several pre-processing steps need to be conducted. This is encapsulated in the diagrams shown below. This was the initial data pre-processing pipeline. Here, hashtag splitting was replaced by stemming. This was due to issues with irrelevant TF-IDF scores [15] assigned to split words in a tweet. On conducting further analysis,

it was found that each of these steps contributed positively to the evaluation results of the algorithms. For instance, stopword removal improved the silhouette score by 15% and reduced clustering time by around 28% for the k-means algorithm and 15% for spectral clustering. All of these preprocessing techniques were performed using the NLTK and Regex libraries in Python.

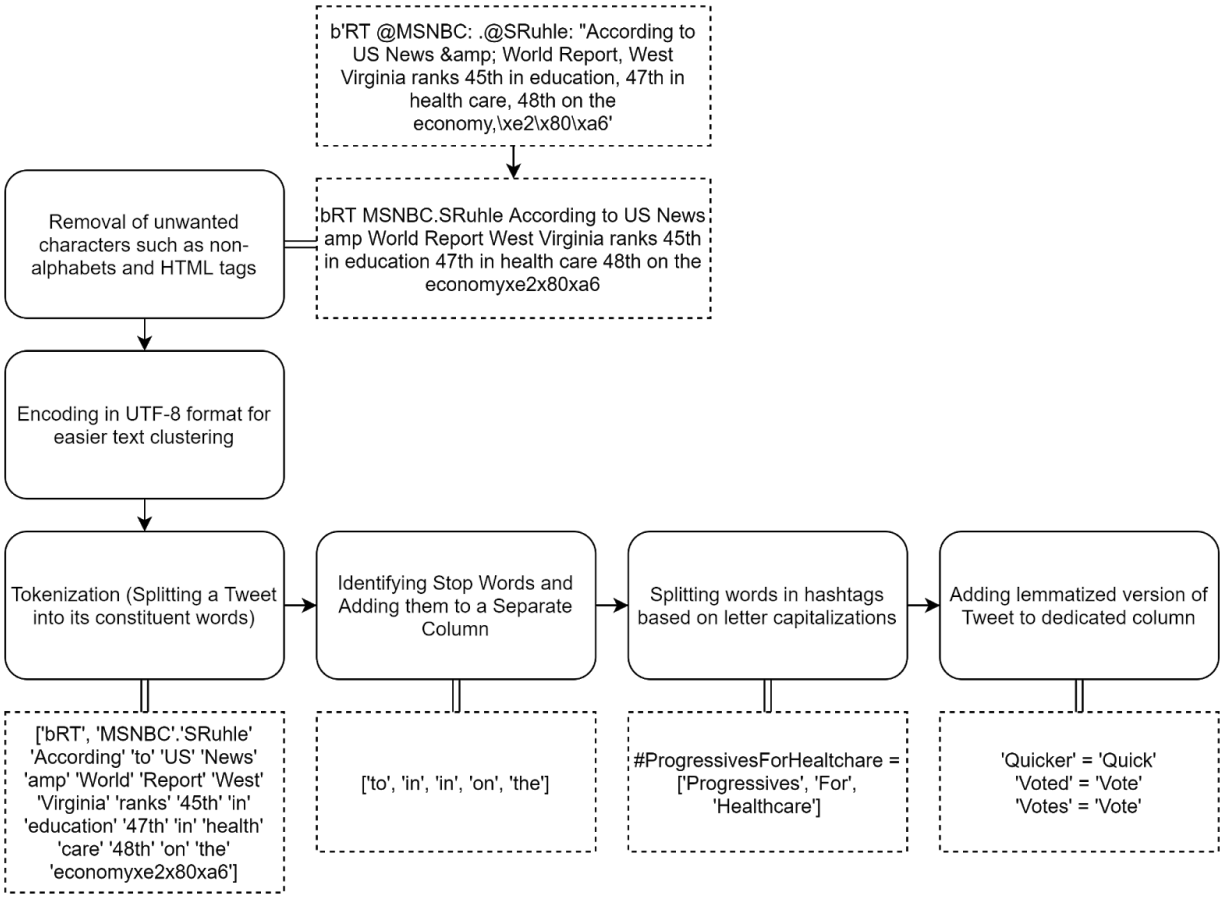



Figure 2 - Standard Data Pre-Processing Pipeline

## 3.3 k-means Clustering with SVD

### 3.3.1 The algorithm

k-means is a popular heuristic algorithm that does not converge to a global optimum. It is sensitive to initialization. In this scenario, the input to the k-means algorithm is the feature vector produced post-TF-IDF score computation and feature vector creation. The objective function can be defined as:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

where *($x_1$ to $x_n$)* are d-dimensional observations, '*n*' is the number of observations, '*k*' is the number of clusters, and '*S*' is a k-dimensional matrix of the within-cluster sum of squares. The algorithm can be summarized in 4 steps - *(1) Randomly choose 'k' cluster centers. (2) Calculate the distance of each point to each cluster center. (3) Assign each point to the nearest cluster center. (4) Recompute the cluster centers based on point assignments. Iterate until no points change their clusters.*

The value of *'k'* can either be 9 (number of political classes in [3]) or determined algorithmically by plotting an elbow diagram and identifying the inflection point. In my case, I found that the ideal number of clusters was around 8. Thus, I decided to go ahead with the Piper Gradient SME typology as detailed in the appendix. The algorithm was implemented using Scikit Learn's inbuilt k-means function.

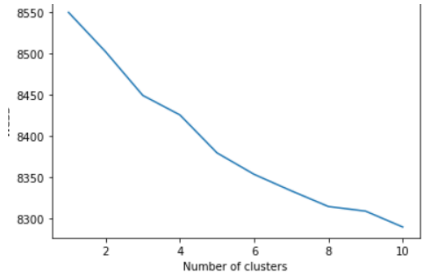



Figure 3 - Elbow Diagram for the ideal number of clusters

### 3.3.2 Results

**Quantitative Metrics:**

| | **Homogeneity Score** | **Completeness Score** | **V-measure** | **Silhouette Score** | **Time to Run (in seconds)** |
|---|---|---|---|---|---|
| **User-Description -based labels** | **0.005** | **0.004** | **0.005** | **0.4170** | **32.7** |
| **Crowd-sourced labels (k = 9)** | 0.028 | 0.027 | 0.028 | **0.3879** | **3.2** |

*(See figure (b) in appendix for Silhouette Score and clustering distribution for k = 8)*

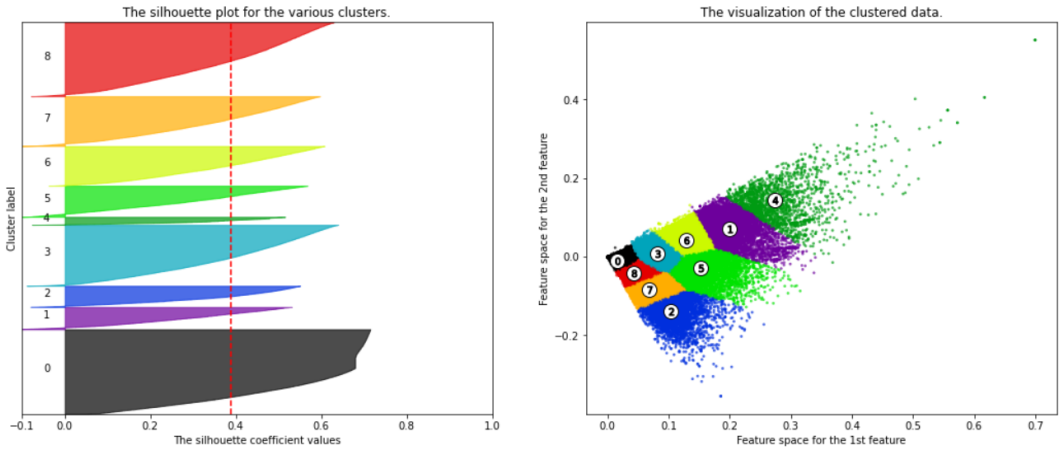



Figure 4 - Silhouette Coefficient by Cluster (Left), Feature Space for 1st Feature (Right)

**Pie Charts:**

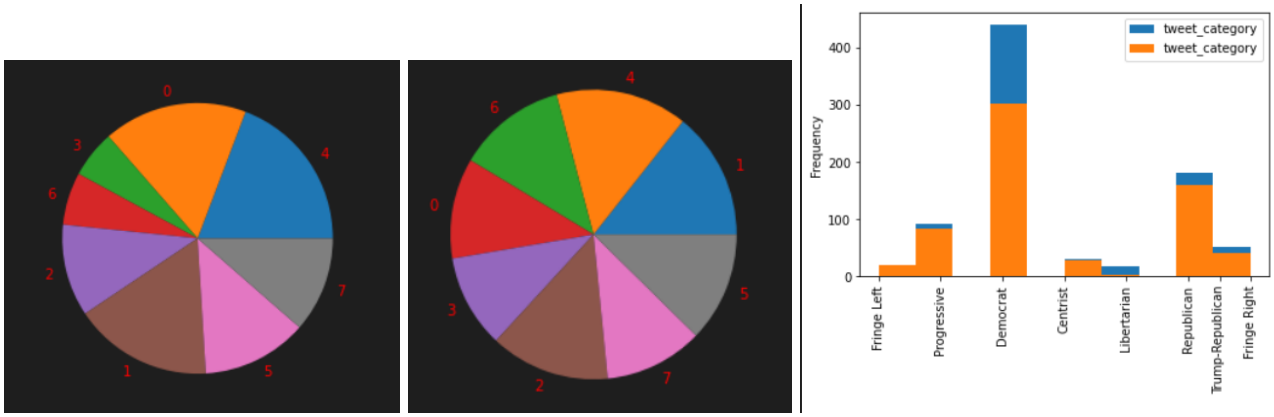



Figure 5 - Distribution of liberals in clusters (left), distribution of conservatives (center), distribution of positive and negative tweets by user category **(orange = negative, blue = positive)**. We see that there is an almost uniform distribution of liberals and conservatives in each cluster.

**Word Clouds:**

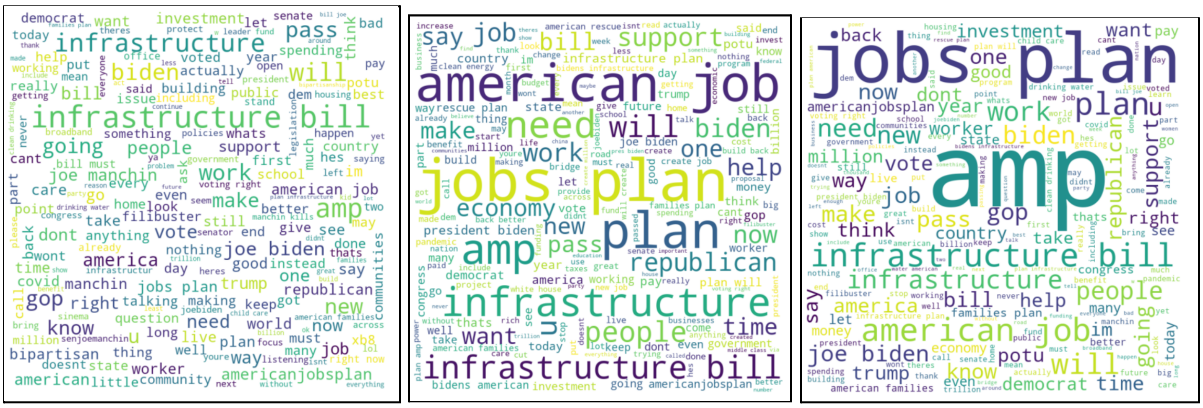


Figure 6 - Standard Data Pre-Processing Pipeline

## 3.4 Mini Batch k-means with Rich Feature Vector

### *3.4.1 The algorithm*

In the normal k-means algorithm, we lose some information due to a compulsory SVD step. This is because all the data cannot be loaded in the memory concurrently without cutting down the feature vector. This is not the case with Mini Batch k-means. Here, we have a similar objective function and algorithmic approach. However, we load a small batch of random data in each iteration instead of the entire set. A convex combination of the prototype values and data is used to perform the clustering. This means that we can feed the entire TF-IDF feature vector without any truncation.

### *3.4.2 Results*

| | **Homogeneity Score** | **Completeness** | **V-measure** | **Silhouette Score** | **Time to Run (in seconds)** |
|---|---|---|---|---|---|
| **User Description-based labels (k = 8)** | 0.044 | 0.039 | 0.041 | **0.4575** | **28.3 (large dataset) 2.3 (medium)** |

| | | | | | |
|---|---|---|---|---|---|
| **Crowd-sourced labels (k = 9)** | **0.098** | 0.066 | **0.079** | **0.4609** | **1.1** |

(See figure (a) in appendix for Silhouette Score and clustering distribution in two dimensions)

**Sample Word Clouds:**

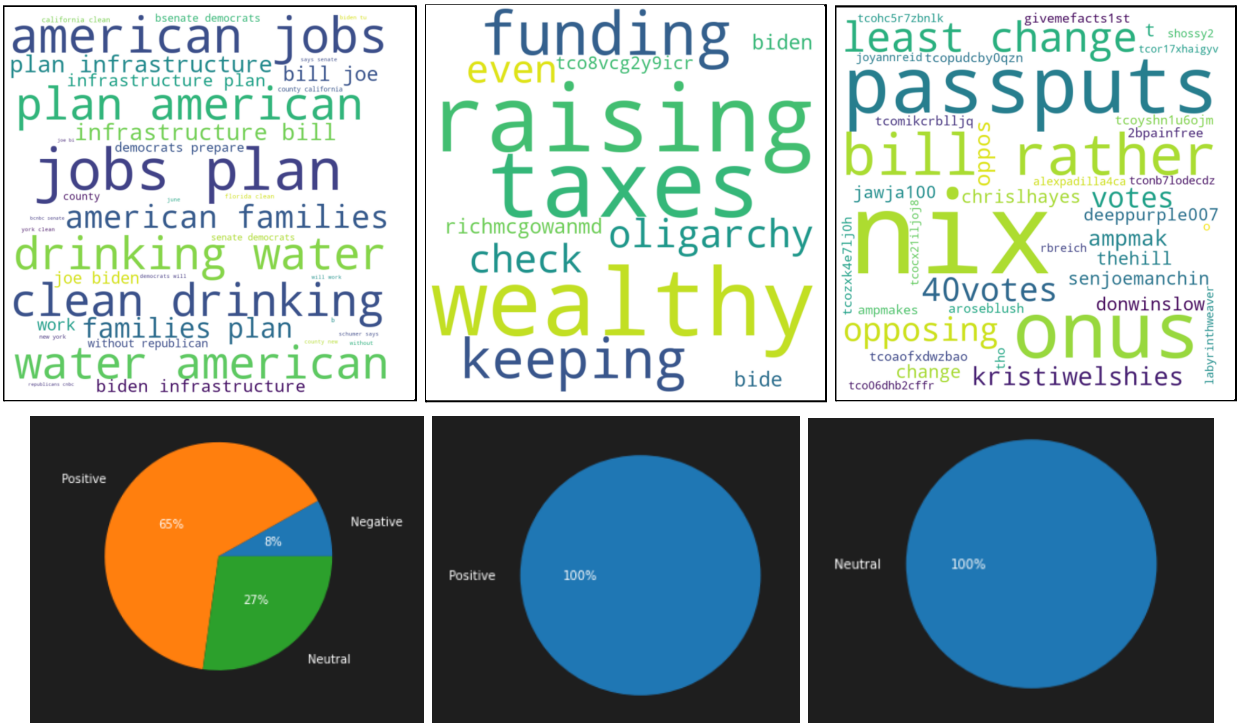



Figure 5 - Generally positive (left-most), highly positive (center), highly neutral (right-most)

**Pie Charts:**

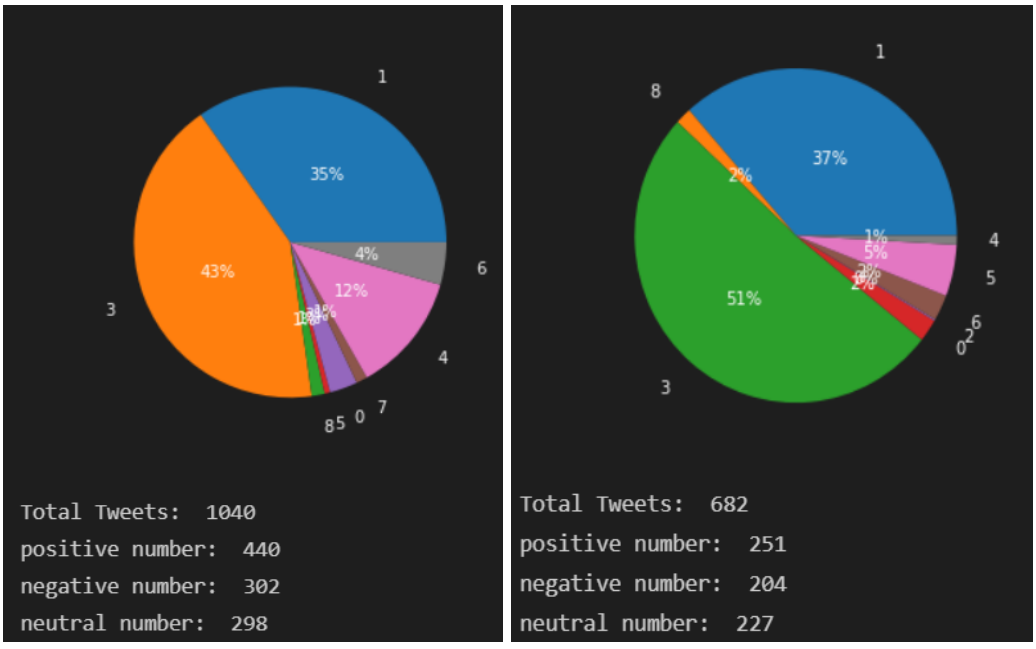



Figure 6 - **Democratic Sentiment (Left)** vs. **Republican Sentiment (Right)** from a random sample of tweets based on cluster numbers

## 3.5 Spectral Clustering

### *3.5.1 The algorithm*

Spectral cluster is a graph-based clustering algorithm that uses similarity matrices to perform clustering after dimensionality reduction. The TF-IDF score acts as a proxy for the similarity metric between words within the graph. Given the similarity matrix, we compute the corresponding **Laplacian matrix (L)** by subtracting it from a diagonal matrix (D). We then take the first 'k' eigenvectors ('k' lowest eigenvalues of L). The p'th row of the resulting matrix of the first 'k' eigenvectors represents the p'th feature of the node in the graph. Now, the nodes in the graph can be clustered using any distance-based clustering algorithm. In this case, k-means is used alongside Dijkstra's algorithm to calculate the distance between nodes.

$D_{ii} = \sum_j A_{ij}.$ , where D is the diagonal matrix and A is the similarity matrix containing the features.

$L^{\text{norm}} := I - D^{-1/2} A D^{-1/2}$ (source)

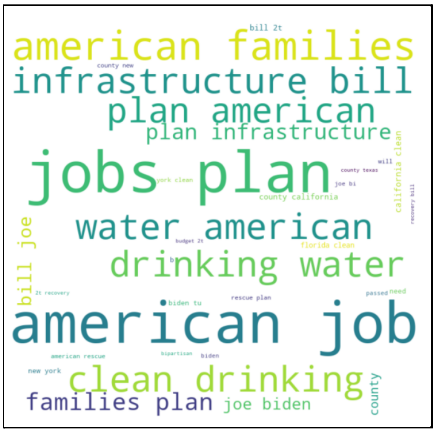
### *3.5.2 Results*

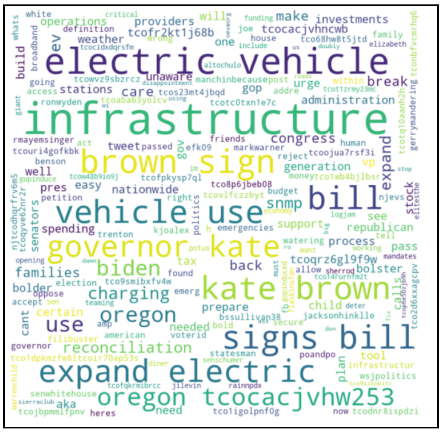
**Quantitative Metrics:**

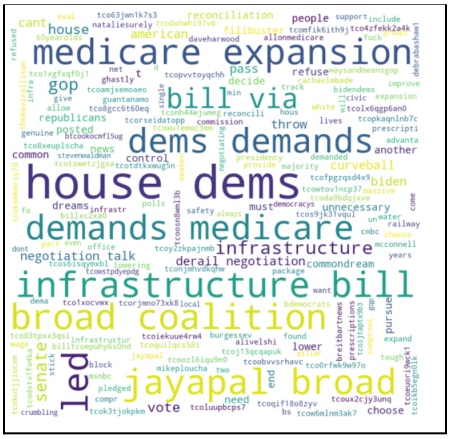
| | Homogeneity Score | Completeness Score | V-measure | Time to Run (seconds) |
|---|---|---|---|---|
| **User Description-based labels** | **0.048** | **0.044** | **0.046** | **105.5**(large dataset)<br>**14.3** (medium) |
| **Crowd-sourced labels (k = 9)** | **0.052** | **0.050** | **0.051** | **1.3** |
| **k-means labels** | **0.216** | **0.213** | **0.215** | **-** |

Note: The k-means labels are based on previous clustering results.

**Word Clouds:**



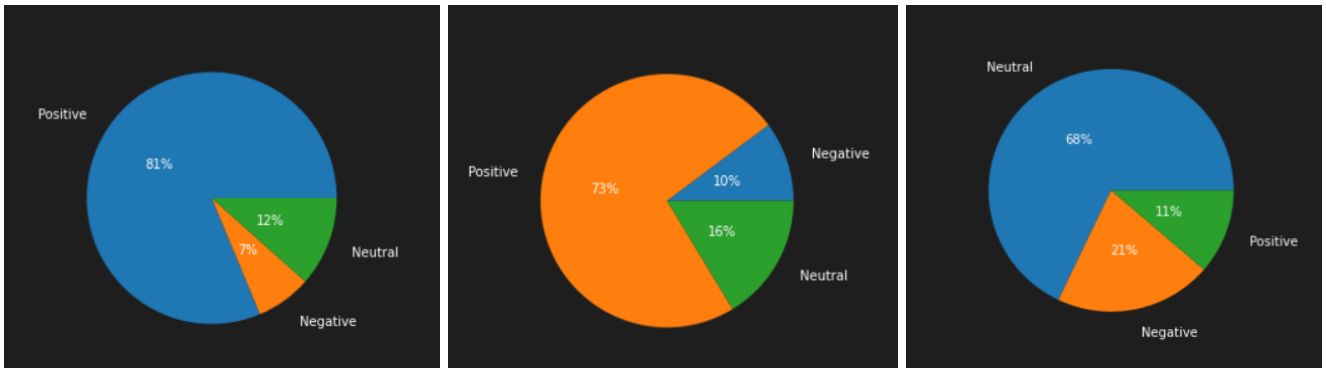



Figure 5 - **Highly Positive (left-most), highly positive (center), dominantly neutral (right-most)**

### 3.6 Comparison and Discussion of Results

The vanilla k-means algorithm distributed liberal groups and conservative groups almost evenly in each cluster (based on both typologies -> greater than 0 = conservative and lesser than 0 = liberal). This meant that there was no political uniqueness or novelty to each group. Thus, it might be possible to conclude that the clustering algorithm is unable to pick up nuances in political ideology through a tweet clustering exercise. To verify this, I performed a hypothesis test at a positivity threshold of 75% (at least 75% of the tweets in the cluster should have the same sentiment) after sampling 50 sets of 100 tweets from the test set. I was able to reject the null at the 0.05 level, given that the null hypothesis was: *the clustering algorithm is capable of picking up semantic and sentimental nuances that relate to political ideology*.

The same hypothesis test **did not** produce similar results with spectral clustering and mini-batch k-means. As seen in the word clouds and pie charts shown above, some of the clusters were indicative of political ideology. They also exhibited uniformity in intra-cluster sentiment. This means that labeling can be done for certain clusters after subjective examination of the word cloud for the cluster. If the word cloud is not sufficiently convincing, the person performing the labeling exercise might have to look at the user descriptions and tweets for the tweets within that particular cluster. Batch K-means was significantly faster than vanilla K-means and also exhibited a better homogeneity score, completeness score, and V-measure. The time to run in terms of seconds might vary based on the running environment, but the difference in time remains largely consistent as a percent value. Spectral clustering takes the longest amount of time to complete, possibly due to the number of intermediate steps involved. Although the metrics for spectral clustering (homogeneity, completeness, V-measure) are similar to mini-batch K-means, every cluster has a sizable number of elements. Mini batch K-means without SVD sometimes results in clusters containing very few tweets, but this is not the case with spectral clustering.

## 4. Ethics Report

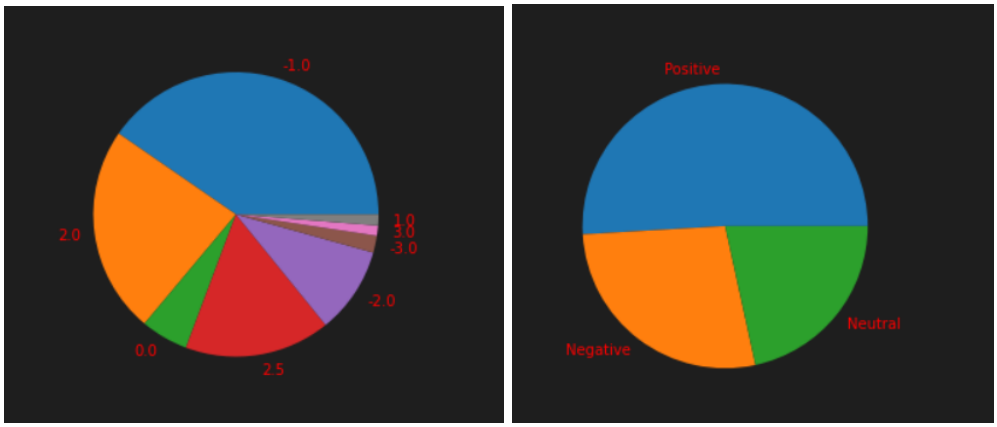
While clustering may seem like a harmless piece of the puzzle in reverse-engineering social media echo chambers, it can have a serious negative impact. As detailed in [17], all AI should be responsible, equitable, traceable, reliable, and governable. Some of these core principles could potentially be violated if clustering results are taken at face value. Throughout my analysis, I found that the initial conditions and parameters could be manipulated for a dataset to obtain a cluster that contained meaningful mapping to political typology. However, the same preprocessing cannot be generalized across topics, making this an unethical approach to

solving the problem. Instead, I believe the purpose of the clustering exercise should be to explore the data and attempt to find a link between the algorithm and the typology ONLY if there is sufficient proof that such a link exists. If we fail to find the proof, we should refrain from making declarative statements, and use the clusters as guidelines for manual labeling. Moreover, we should examine the desirability of such an echo chamber reverse-engineering exercise for people across the political spectrum. The result of the exercise should not be a force-feeding system where people are constantly forced into being exposed to tweets, posts, or videos from different clusters against their will.

**1) What data biases have been discovered?**

An initial exploration of the data immediately revealed that a disproportionate number of people support the Infrastructure Bill in the provided dataset. This was verified through label assignment based on user description as well as sentiment analysis. Using the Piper SME typology, the number of individuals who fall into the '-1.0 (Democrat)' category is far more than any other group. The number of positive tweets overall in the large dataset is more than half the total number of tweets.

From a broader viewpoint, the data is composed of short tweets and user descriptions that are barely over 400 characters when combined. As one of the early manual labeling exercises demonstrated, placing tweets in one of 9 categories turned out to be a complicated affair for a human. Expecting a clustering algorithm to extract political leanings from such short pieces of text may be unrealistic. The fact that not all the users in the dataset have meaningful user descriptions further exacerbates the problem.




Lastly, the data is completely focused on a single political issue - the Infrastructure Bill. Using user descriptions to intuit their political leanings may not be suitable in all cases since people are likely to change their political stances based on issues. The 2017 Pew Research study [3] itself talks about how people vote on issues, and not always past loyalties. Socio-economic conditions are important factors that are fundamentally dynamic. Thus, there is an acute data diversity problem with the given dataset. This can be remedied by providing more diverse data across different social and political issues gathered over a longer period. We can then perform **topic clustering** as shown in [26] to first classify tweets into their respective topic groups before performing a deeper statistical analysis across all groups. *Although I initially intended to perform topic clustering to find subtopics within the infrastructure bill universe of tweets, I was unable to generate meaningful topics that represented the tweets that inhabited the topic clusters.*

**2) What are the appropriate metrics to assess the capability's output? What would be tolerable in the actual operation of the Gradient system?**

The appropriate metrics to assess the capability of the output are chosen to be homogeneity, completeness, V-measure, silhouette score, and rand-index. A comparison of the values of these metrics for the outputs of various algorithms on different datasets is shown in section 3. All of these metrics are chosen since they adequately measure the purity of a cluster. More specifically, homogeneity measures the frequency of co-occurrence of labels. Completeness is similar to homogeneity score but is sensitive to the order in which the parameters are passed. The V-measure is the harmonic mean of homogeneity and completeness. It is given by: (1 + beta) * homogeneity * completeness / (beta * homogeneity + completeness). As shown in [19], this entropy-based evaluation method is independent of the clustering algorithm and dataset, and also accounts for the problem of matching where the clustering is only done on a small portion of the data set. This makes it an ideal candidate to be used in this context.

The Rand Index is given by the ratio: (TP + TN) / (TP + TN + FP + FN), where TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative
The adjusted Rand Index adjusts for the expected value. Therefore, it can range from -1 to +1. A low adj. Rand Index indicates that there are fewer correct labels than expected.

While all these metrics can be powerful indicators, they serve only as guiding figures, and not absolute deciders. As seen in the above analysis, the homogeneity family of scores can be quite low in several cases. This is due to the presence of 9 labels (sometimes 8). It is quite likely that a ‘fringe right’ tweet gets classified into a majority ‘Trump Republican’ cluster. This is why a homogeneity score even close to 0.1 is a rather high value. While none of these scores (except adjusted rand index) should slip below 0.04 for a reasonably clustered dataset, a subjective analysis of the word cloud or contents of the cluster are almost always a necessity. The evaluation metrics can simply serve as cut-off values to assess the success or failure of an algorithm in forming political typology clusters.

**3) How may introducing an accountable human to the Gradient CONOPS produce cognitive biases and/or confirmation bias?**

In [27], we find that there are three types of bias that can affect any computer system - Pre-Existing Bias, Technical Bias, and Emergent Bias. The ‘individual’ component of the pre-existing bias will naturally play a strong part in the Gradient CONOPS. In [28], Joyce Ehrlinger et al. talk about how people tend to frequently settle for ‘good enough’. In the context of clustering, examining a word cloud that may seem to be in line with the sentiment analysis can lead to premature labeling of the cluster. Although not immediately apparent, we also suffer from societal bias. We performed the entire clustering exercise while only discussing potential ethical challenges only within a closed group of people who collectively believe (to some extent) that social media echo chambers are a problem and that machine learning is a reasonable approach to solve the problem. While I may still believe in this proposition personally, my opinions are irrelevant in the larger context. Thus, an entire project team may potentially have fallen prey to some form of cognitive bias during an ethics discussion.

There are several biases present on a technical plane as well. There is a computer tech bias since we have decided to use clustering as the approach, while we could have explored other ways of doing the task, such as decision trees. There is a decontextualized algorithm bias

since we are only focusing on tweets and user descriptions to make a rather huge judgment on the user's political ideology. We are also only focussing on a single dataset that represents just one political and social issue. There is very little data diversity. Finally, we are introducing codifying abstract human concept bias by oversimplifying tweet sentiments. We do not account for sarcasm, quotes, or trolls. One could argue that random number generation bias is also present since the algorithm initializations and hypothesis tests rely on random number generation.

**4) How should undesired bias and potential impacts of the bias be communicated to a Gradient user and administrator who interact with the technology and output data?**

A Gradient user or administrator must be made explicitly aware that the clustering exercise can sometimes result in clusters that are ripe for misinterpretation. Gradient administrators should be encouraged to actively explore word clouds, frequently occurring terms, and the actual rows in the dataset that correspond to the elements in a particular cluster. Moreover, it should be made clear that the clusters do not always result in good mapping to political typology. In fact, it is more likely that a cluster of tweets made in 2021 on an infrastructure bill corresponds entirely to a political typology based on socio-economic factors in 2017. While the user or administrator may be tempted to leap to a conclusion based on the results of clustering, they must be encouraged to experiment with different test sets, initial conditions, parameters, and even typologies before coming to any conclusions.

**5) If analyses rely on third-party technology, software, or data product, what additional risks may be associated with that third party's assumptions, motives, or methodologies?**

Inspired by Faruqui et al. [22] on the problems with word embeddings, I picked a more neutral TF-IDF score for vectorization of the data. In the paper, they mention how word embeddings can be extremely inaccurate. Word embeddings can also be downright discriminatory at times, even encouraging negative social associations related to homophobia and racism as explained in Oscar Schwartz's 2019 article [30] in IEEE Spectrum. In the case of the Twitter dataset, due to several proper nouns, bad spelling, and the presence of hashtags in tweets, many of the terms do not even have a vector representation in libraries like word2vec and GloVe [21]. They end up being assigned garbage values that interfere with the whole process.

There are also issues with the sentiment analysis function in the NLTK Vader package as demonstrated in [23]. Its inability to perform well in different domains, inadequate accuracy and performance in sentiment analysis based on insufficient labeled data and incapability to deal with complex sentences are some of the primary risks. Using sentiment analysis as a way of identifying support for the bill might by itself be dubious as shown in the 'negative' sentiment examples shown below:

- *"Can't believe what America has come to: \nIntegrity is now considered boring and not newsworthy. \nWhy doesn't the media talk\xe2\x80\xa6""*
- *"#endthefilibuster #endworkingwithtrumplicans pass the infrastructurebill as is say no to moscowmitch"*

Both of these tweets were classified as ‘negative’. However, it can clearly be seen that they are either neutral or in support of the bill. Therefore, it is important to use semantic analysis to see if clusters are homogenous, and not to directly assign labels post clustering.

In Lindsey Andersen’s article [29], we see that there is a violation of the principles of transparency and explainability with the Vader package. It acts more like a black box in this scenario. To remedy this, we should never shift the direct responsibility of labeling entirely to AI. Instead, the task should be cognitively shared between the AI and the human performing the labeling.

**6) Given the purpose of the AI, what level of explainability or interpretability should be required for how the AI made its determination? For third-party dependencies, how can you ensure a level of explainability or interpretability?**

The interpretability of my clustering exercise can be broadly divided into four categories:

1) Why was a particular data preprocessing step carried out? Does it obfuscate or morph the data in any undesirable fashion?
2) Why was the null hypothesis (shown in section 3.6) rejected or not rejected for a particular clustering exercise?
3) What features are being used for clustering?
4) How were the sentiment labels assigned in the first place?

1 - The first two questions are purely related to the clustering pipeline without any dependency on contentious packages such as the NLTK Vader package. The Scipy implementation of k-means or spectral clustering is a fairly mathematical process that does not suffer any subjective interpretability or explainability issues. To address the first question, none of the preprocessing steps are opaque. Each of the steps contributes positively to the clustering metrics (see figure(e) in the appendix). While some syntactic aspects of the data may be lost, much of the morphological and semantic aspects of the data are preserved.

2 - To address the second question, the user or administrator should be aware that the test is being carried out at an alpha level of 0.05. The test statistic being used is the average positivity rate post sentiment analysis of all the tweets within a particular cluster. The threshold for rejection (alpha-level) and the threshold from which the mean value of positivity is subtracted can be selected by the administrator or user.

3 - In the case of the mini-batch k-means, the entire feature vector is fed into the model. Thus, all of the words in a particular tweet or user description are used to build the model. In the case of any sort of eigenvector decomposition, principal component analysis, and SVD, only a part of the overall variability in the data remains intact. There is a certain degree of obfuscation, and there is no tangible interpretation for the feature vector in terms of the actual words being used for clustering. Unfortunately, there is little we can do to preserve all the information in the original feature vector.

4 - As explained in [24] and the official NLTK documentation, we see that tokenized words are used as simple unigrams for sentiment analysis. Scores are assigned based on the independent semantics of each word in a tweet. This ignores word relationships and words that are not present in the corpus (especially hashtags). The linear combination of scores gives the

sentiment score for the tweet. Thus, it is clear that biases are present at every step of the process. To maintain a reasonable level of interpretability, the sentiment analysis scores of each of the constituent words of a tweet can be examined. We can subjectively see if these scores make sense in the context of the tweet and the user description.

Most of the third-party dependencies, in this case, are all open-source and easily accessible.

**7) What processes need to be in place during operation for discovering undesired bias and drawing conclusions?**

Four broad processes can be put in place to curb the effects of undesired biases:

1) If a strong positive, negative, or neutral sentiment is exhibited by a cluster, observe the number of elements within the cluster. If the number of elements is far too less, we cannot label the cluster without experimenting with more data.

2) If a clustering exercise results in a low homogeneity score, it does not mean that it should be immediately cast out. Carrying out a thorough (and time-consuming) subjective analysis of the cluster can tell us if there is actually any political novelty to the cluster. If the homogeneity score is low, but the cluster does turn out to be politically relevant, we might have to reconsider the approach to assigning proxy labels and evaluating the algorithm for that particular dataset.

3) In her article on the ethical pitfalls of AI [29], Lindsey Andersen lays out seven principles for the responsible use of AI. In line with the principles of Human Computer Interaction, the clustering exercise should also be designed with the end-user in mind. In her second point, Andersen mentions that we need to understand the ecosystem. This is perhaps the most crucial step to complete before conclusions can be drawn. When we are actively dealing with a highly political environment, there is very little room for personal biases to creep in. Instead of one accountable non-biased human investigating the clusters, it might be a better idea for a group of humans with self-declared political biases to explore the semantics of each cluster and come to a conclusion.

4) Another important aspect of Andersen's list is the use of open standards, open data, open innovation, and open source. Throughout the process of clustering, an attempt must be made to use libraries that are freely available on the internet such as Scipy. This ensures that all parties can examine the source code. Moreover, internal verification standards can be put in place to ensure that open source tools and data are being used to build the models at all times.

**8) Is the entire project and analysis constructed on a stable base of assumptions? Are social media echo chambers really as big a problem as we make them out to be?**

While this question was not part of the list of questions to be addressed in the ethics report, I felt it was important to include this. This will be a more subjective and less formal exploration based on personal opinions tied to the results of this analysis. While several papers such as [1] and [2] show the existence and dangers of social media echo chambers, it is

important to ask if the gravity of the situation is truly as dire as it may seem. In [16], Elizabeth Dubois & Grant Blank argue that there is a sizable portion of the population that does not subscribe to social media echo chambers, despite actively using social media. The paper goes on to argue that the horrors of partisan segregation have been largely exaggerated.

The authors of [16] acknowledge that single-media studies have been successful in demonstrating the presence of social media echo chambers. However, there is very little research on echo chambers across media for the same group of users. For instance, there is insufficient research to indicate that people who belong to a certain group on Facebook necessarily belong to the same kind of group on Twitter. They may be using the two platforms for entirely different reasons. In [25], Henry Lothane from the Icahn School of Medicine talks about how influential individuals have influenced society since time immemorial. Fascist regimes in the past rose without any need for social media. Through large-scale propaganda and misinformation, they convinced entire populations to stay silent in the face of tyranny, particularly during the second world war. Perhaps, in some ways, the diverse opinions on social media are acting as a means of stopping certain groups from exerting undue control over their populace. In fact, in modern societies where social media is heavily regulated, we see forced echo chambers that are far worse than the echo chambers that naturally form on open social media. This is not to conclusively say that echo chambers are not inherently detrimental to society; it is an alternate avenue of thought that may require further exploration and research.

# 5. Conclusions and Future Work

In this project, I was able to compare different text preprocessing and clustering techniques. I found that some techniques such as the vanilla K-means algorithm with support vector decomposition were completely ineffective at capturing political nuances. I would recommend the usage of non-truncated feature vectors, especially in the case of the k-means algorithm with short textual data (tweets). Moreover, algorithms such as mini-batch k-means and spectral clustering have the potential to assist humans in labeling clusters based on political typology. Implementing and comparing other algorithms such as K-medoids and Louvain may be worthwhile as future endeavors.

In terms of ethical considerations, I am particularly interested in exploring the formation of social media echo chambers. It might also be fruitful to perform an investigation of ideological radicalization in terms of psychology and human history. While there seems to be ample correlation established between social media usage and radicalization, the counterfactual is not sufficiently explored. Is social media simply a new means of using the radicalization playbook that was exploited by revolutionaries and leaders of the past? Is social media also acting as a deterrent to rampant polarization? The answers to these questions might take us on intriguing avenues.

# 6. References


[1] Bright, J., 2016. Explaining the Emergence of Echo Chambers on Social Media: The Role of Ideology and Extremism. SSRN Electronic Journal.

[2] Boutyline, A. and Willer, R., 2016. The Social Structure of Political Echo Chambers: Variation in Ideological Homophily in Online Networks. Political Psychology, 38(3), pp.551-569.

[3] Pew Research Center - Political Typology Reveals Deep Fissures on the Right and Left - https://www.pewresearch.org/politics/2017/10/24/political-typology-reveals-deep-fissures-on-the-right-and-left/

[4] Harakawa, Ryosuke & Takimura, Shoji & Ogawa, Takahiro & Haseyama, Miki & Iwahashi, Masahiro. (2019). Consensus Clustering of Tweet Networks via Semantic and Sentiment Similarity Estimation. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2936404.

[5] R. Soni and K. J. Mathai, “Improved twitter sentiment prediction through cluster-then-predict model,” Computing Research Repository, vol. abs/1509.02437, pp. 1–5, 2015 (http://arxiv.org/abs/1509.02437).

[6] F. Sebastiani, “Machine learning in automated text categorization,” ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.

[7] M. Vicente, F. Batista, and J. P. Carvalho, “Twitter gender classification using user unstructured information,” in Proc. IEEE Int. Conf. Fuzzy Systems (FUZZ-IEEE), 2015, pp. 1–7.

[8] J. C. B., R. Ehrlich, and W. Full, “Fcm: The fuzzy c-means clustering algorithm,” Computers & Geosciences, vol. 10, no. 2, pp. 191 – 203, 1984

[9] G. Ifrim, B. Shi, and I. Brigadir, “Event detection in twitter using aggressive filtering and hierarchical tweet clustering,” in Proc.SNOWDC@WWW, 2014

[10] J. C. B., R. Ehrlich, and W. Full, “Fcm: The fuzzy c-means clustering algorithm,” Computers & Geosciences, vol. 10, no. 2, pp. 191 – 203, 1984.

[11] S. Miyamoto, S. Suzuki, and S. Takumi, “Clustering in tweets using a fuzzy neighborhood model,” in Proc. IEEE Int. Conf. Fuzzy Systems, 2012, pp. 1–6.

[12] Baillargeon, Sophie & Halle, Simon & Gagné, Christian. (2016). Stream clustering of tweets. 1256-1261. 10.1109/ASONAM.2016.7752399.

[13] C. J. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in Proc. Int. AAAI Conf. Weblogs and Social Media, 2014, pp. 216–225.

[14] FACT SHEET: The American Jobs Plan - WH.GOV (Statements and Releases - March 31, 2021)

[15] Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.

[16] Elizabeth Dubois & Grant Blank. The echo chamber is overstated: the moderating effect of political interest and diverse media. Volume 21, 2018 - Issue 5: Communication, Information Technologies, and Media Sociology (CITAMS) Special Issue. Published online: 29 Jan 2018.

[17] AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense

[18] Crime and Policing - Mark H. Moore, Robert C. Trojanowicz, and George L. Kelling

[19] Rosenberg, Andrew & Hirschberg, Julia. (2007). V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure.. 410-420.

[20] Efficient Estimation of Word Representations in Vector Space by Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, arXiv:1301.3781 (cs). Submitted on 16 Jan 2013 (v1), last revised 7 Sep 2013 (this version, v3).

[21] GloVe: Global Vectors for Word Representation by Jeffrey Pennington, Richard Socher, Christopher D. Manning. Computer Science Department, Stanford University, Stanford, CA 94305.

[22] Faruqui, Manaal, Yulia Tsvetkov, Pushpendre Rastogi and Chris Dyer. “Problems With Evaluation of Word Embeddings Using Word Similarity Tasks.” ArXiv abs/1605.02276 (2016): n. pag.

[23] Shahnawaz and P. Astya, "Sentiment analysis: Approaches and open issues," 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 154-158, doi: 10.1109/CCAA.2017.8229791.

[24] Hutto, C.J. & Gilbert, Eric. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

[25] Lothane, Henry. (2006). Mass psychology of the led and the leaders. International Forum of Psychoanalysis. 15. 183-192. 10.1080/08037060600924983.

[26] Llewellyn, Clare & Grover, Claire & Oberlander, Jon. (2016). Improving Topic Model Clustering of Newspaper Comments for Summarisation. 43-50. 10.18653/v1/P16-3007.

[27] Three kinds of bias in computer systems - THE INSTITUTE OF TECHNOLOGICAL ETHICS

[28] Ehrlinger, Joyce & Readinger, W.O. & Kim, Bora. (2016). Decision-Making and Cognitive Biases. Encyclopedia of Mental Health. 10.1016/B978-0-12-397045-9.00206-8.

[29] Artificial Intelligence in International Development: Avoiding Ethical Pitfalls by Lindsey Andersen

[30] In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation By Oscar Schwartz

# 7. Appendix

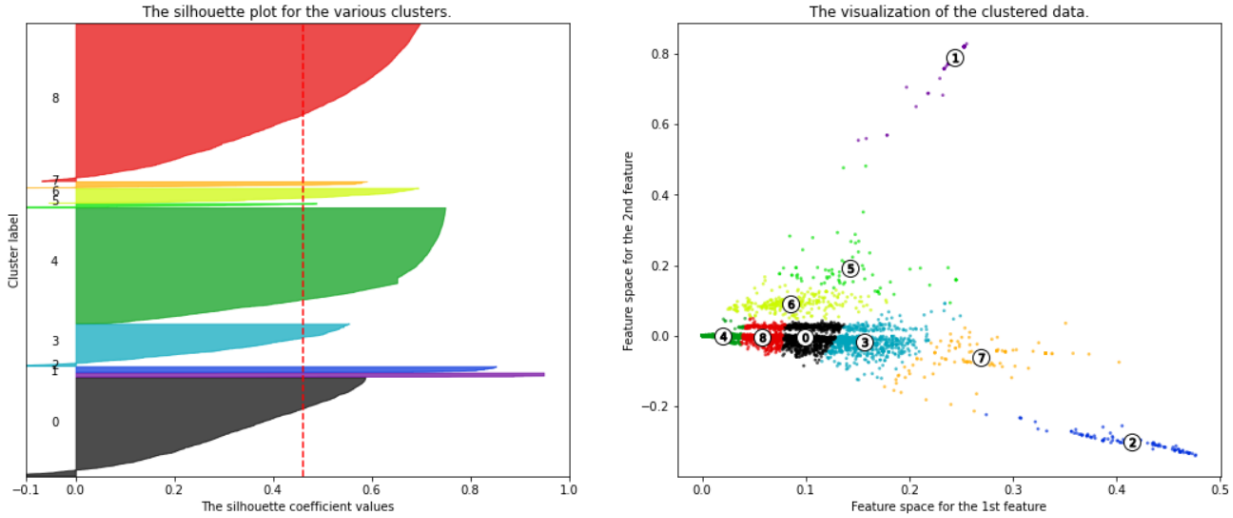



Figure a - Silhouette scores and visualized clusters for Mini Batch k-means on Medium Dataset

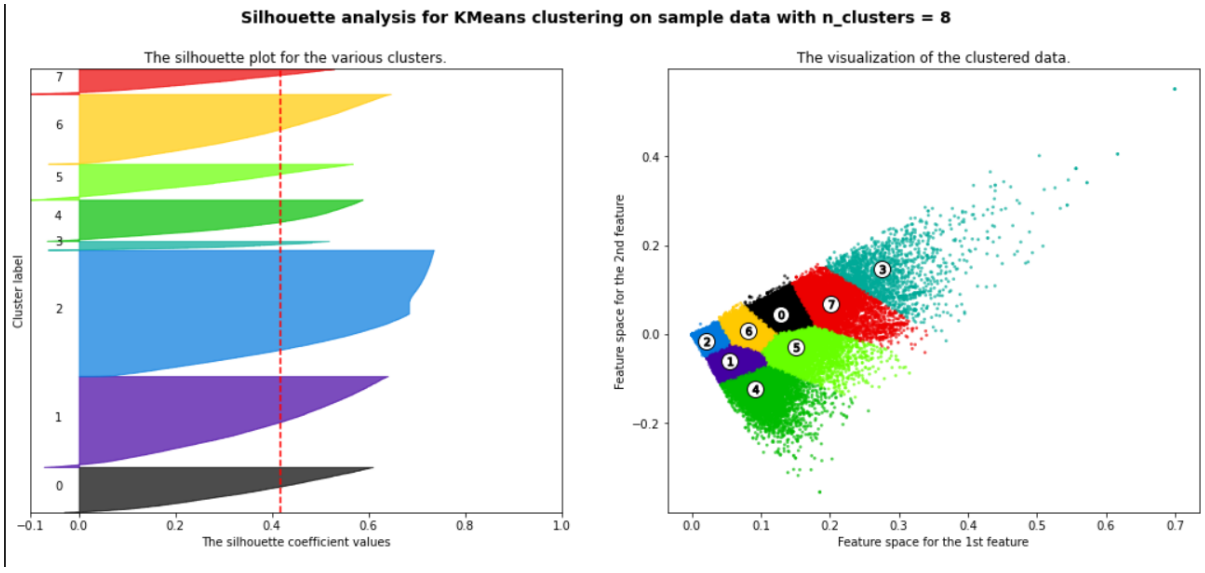



Figure b - Silhouette scores and visualized clusters for k-means on Large Dataset for k = 8

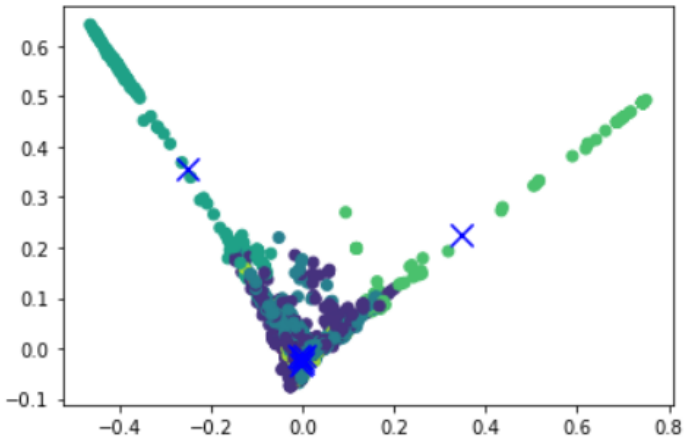



Figure c - Cluster centers for Mini-Batch k-means

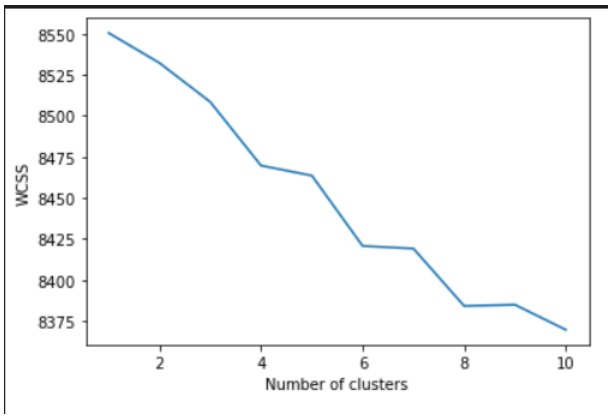



Figure d - Mini-Batch k-means elbow diagram

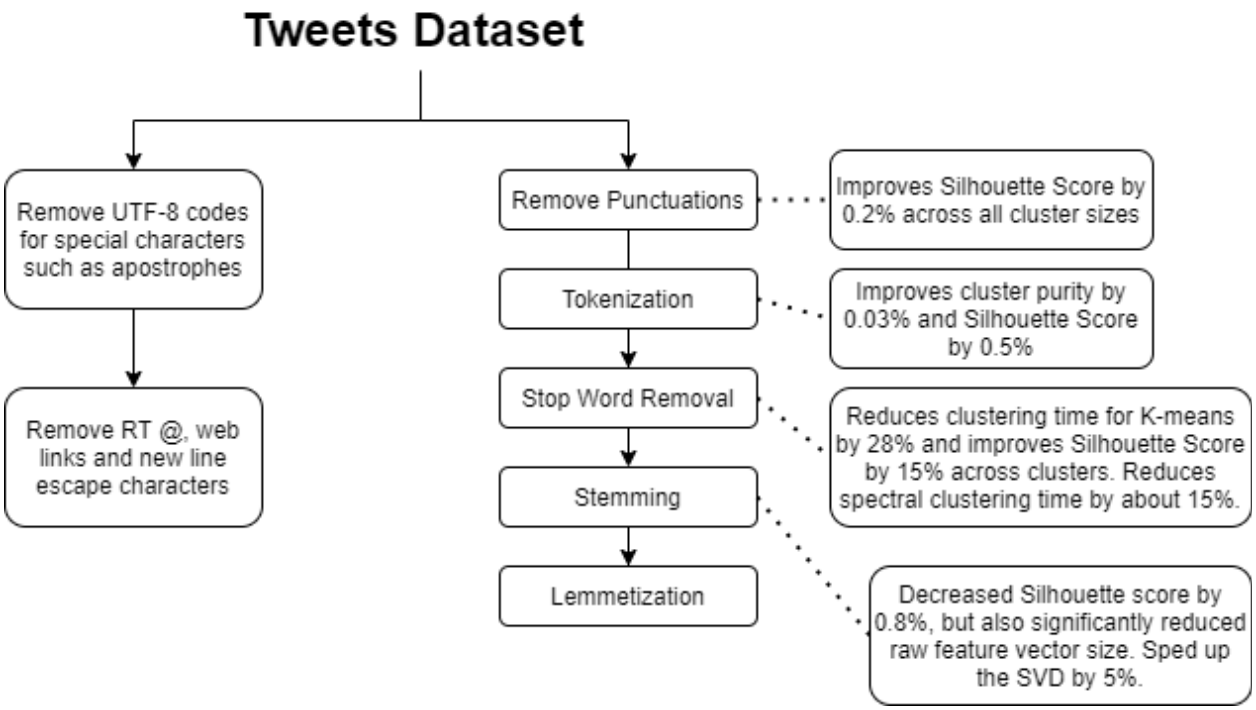



Figure e - Contribution of each preprocessing step

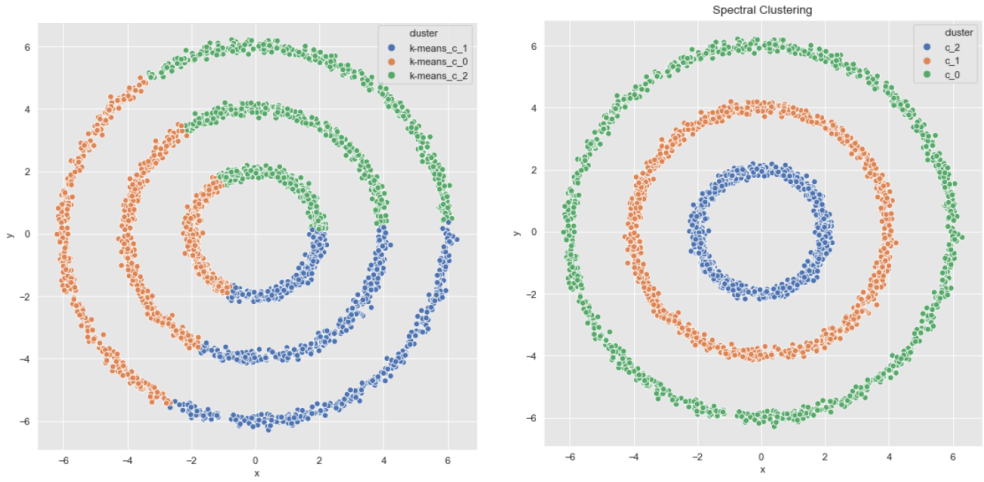



Figure f - k-means (left) vs. Spectral Clustering (right). Source: Getting Started with Spectral Clustering by Dr. Juan Camilo Orduz