The Making of a Controversial Issue:

A Multi-level Approach to Quantify Polarization for a Topic

Allen Habibovic
Northeastern Illinois University
Primary Academic Advisor: Professor Marcelo
Co-Advisor: Professor Iacobelli

Table of Contents

1. Abstract

Several studies have been conducted to better understand, analyze, and quantify social polarization, primarily in political domains through graph-based techniques. Only a few recent research papers have proposed a single metric for polarization. The goal of this research project is to understand the underlying mechanics of social polarization and to be able to quantify a measure of polarization for a particular topic with non graph-based methods. Here we report a procedure that analyzes the distribution of six different properties that define a polarized topic. The polarized distribution is constructed by first performing stance detection, utilizing Twitter data for the BeRT model. Finally, we apply statistical calculations on the distribution and combine these measures into a single metric for polarization.

2. Introduction

We begin by properly defining social polarization. Social Polarization occurs in a topic that generates large amounts of tweets and segregates into two or more clearly opposing sub-groups with conflicting beliefs. This is typically seen in domains such as politics, gender, sports, racial, and virtually any controversial topic. Also these segregated sub-groups usually contain biased or emotional beliefs with very positive sentiments or negative sentiments overall. Throughout this project, a Natural Language Processing(NLP) framework and procedure is provided to uncover the underlying mechanics of social polarization in prevalent and relevant topics utilizing twitter data, and to be able to quantify a single unit of measurement. The field of NLP is considered a difficult problem in computer science. The nature of human language and cultural slang seen on twitter makes it increasingly difficult for computers and artificial intelligence models to understand and process text and perform specific tasks. However, recently with the advent of the Transformer Architecture, more problems are being solved with state of the art results and performance measures. I implement a specific type of transformer language model, BeRT(Bidirectional Encoder Representations from Transformers) for the task of binary text classification for stance detection. The purpose of performing stance detection of tweets for a topic is to develop and further analyze the belief distribution. This is the core component of this research project that differentiates it from other studies. The idea is to calculate and measure six defined properties of the distribution which are inherent for polarized topics and to combine these measures into a single unit of measurement. This gives us a way to rank, compare and contrast topics under analysis.

3. Background Information/Related Works

Most of the previous works for identifying and quantifying polarization are all observed from social media platforms usually in the field of politics, and predominantly the proposed measures of polarization are graph-based techniques.

An early effort in measuring polarization was done in 2013 by *Calais Guerra*. [3] In this study, the researchers again collected tremendous amounts of data from Twitter on a multitude of topics. The main idea was to construct graphs for topics to determine if the graphs can be segregated into two opposing belief groups. The author used the concept of modularity to measure this level separation in the graphs. Modularity is a simple graph calculation where the proportion of the edges that connect to the belief groups minus the expected fraction if the edges were distributed at random. For example, the higher the calculated modularity rate between the segregated belief groups the higher their polarization. However the authors concluded that modularity is an indirect measure for polarization and considered proposing a new measurement that is based on the topics and belief groups boundaries. Their polarization metric is defined on the continuous scale [0, 1] where 0 is non-polarizing, and 1 is polarized.

In a 2017 research study written by *Kiran Garimella*, *"Quantifying Controversy on Social Media".* [4] Her approach to quantify controversy or polarization is based on a graph-based three staged pipeline procedure. In the first stage, a conversation graph is created that would represent the activity relative to a single topic of discussion. This graph is a simple data structure with nodes and vertices, that incorporates meaningful connections between twitter users, hashtags, topics, and the posts. In the second state, the conversation graph is inputted into a graph partitioning algorithm to separate into

two groups. The third and final stage is to measure the controversy/polarization. This controversy measure captures how distant the two partitions are apart. This is done by taking the partitioned graphs and running several algorithms such as random walks, clustering algorithms, and lower dimensional embeddings.

4. Methodology

In order to analyze a topic's polarization, first we need to construct a comprehensive dataset by collecting tweets using the Twitter API V2. As for an example, we will study how polarized the topic of ‘vaccination’ became as a result of the Covid-19 pandemic. We gather tweets from January 2021 until January 2022 and filter these tweets by hashtags: ‘#covid, #vaccines, #vaccineswork, #vaccinessavelives, #vaccineskill, #vaccinesdontwork, #vaccinessideeffects, …’. Additionally we only gathered english tweets and excluded retweets and replies.

For any NLP framework and machine learning model, we have to follow a series of preprocessing steps to ensure the textual data will be compatible for models. Preprocessing steps taken included: lowercase tweets, remove usernames and urls, remove emojis and punctuations, segment multiword hashtags, and stemming and lemmatization.

Before we can study the distribution of the beliefs or stances of these collected tweets, we need a way to automatically label these tweets as pro-vax(1) or anti-vax(0). Instead of using unsupervised clustering algorithms and document topic modeling methods, logically we can label tweets according to the hashtags they contain. For example, if a tweet contains the hashtag, ‘#vaccineswork’ this would be a pro-vax

stance and it would get labeled with a 1. This type of label propagation was seen in a 2020 paper by *Brian Sharber* who analyzed political polarization from news media websites. He labeled right-wing content from well known right-wing political blogs with a 0, and a 1 from left-wing political blogs. [2]

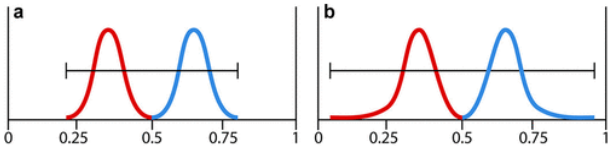
With a labeled dataset, we can now fine-tune a BeRT model for the task of binary text classification. The reason for fine tuning a BeRT model for stance detection purposes is that we can later input new unseen tweets and quickly uncover the stance of that tweet. This also allows us to get continuous values, in the range 0 to 1, for the stance or beliefs instead of a binary quantity.

Now we can study the distribution of these tweets and their respective stance or belief towards the topic. Specifically we look at 6 properties or measures of the distribution that define polarization: spread, dispersion, coverage, distinctness, group consensus, and size of groups. The properties and pictures discussed below were taken from a blog post from the University of Chicago. [1] However the statistical calculations were defined independently of the article.

Property 1: Spread, refers to the total width of the opinions, or how far apart are the extremes in the distribution. This is the range of the opinions, calculated as the difference of the maximum stance value and the minimum stance value.

$$\{STANCES\} = X,\ N = |\{STANCES\}|,\ Xi \in X$$

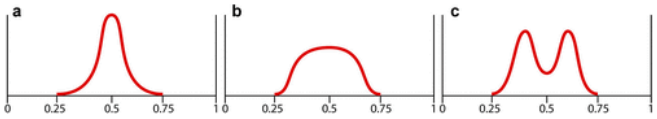
$$Spread = Range = max(X) - min(X)$$



Distribution B is more polarized as it has more spread(range), the difference between the maximum and the minimum belief is greater than that of distribution A.

Property 2: Dispersion, is the statistical dispersion of the distribution such as variation, standard deviation or entropy(disorder). For our purposes we select to do the standard deviation of the distribution. The greater the dispersion, the more polarized the topic is.

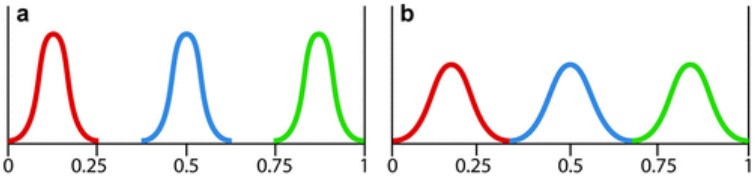
$$Dispersion = \sigma(X) = \frac{\sqrt{\Sigma(Xi - \overline{X})}}{N}$$



Distribution c shows greater polarization in the sense of dispersion than does belief distribution b, which is greater than distribution a.

Property 3: Coverage refers to how polarized belief groups tend to take up a small range of beliefs and center near the median. We can calculate the coverage by taking the number of empty areas in the belief distributions. More empty slots show that the ideas are centering around a central belief.

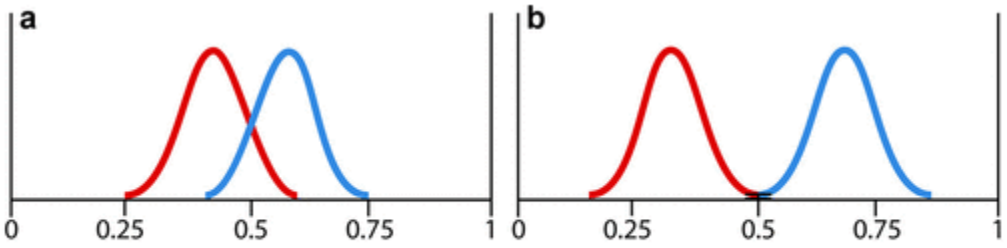
$$Coverage = \%\ of\ belief\ spectrum\ that\ is\ empty$$



Distribution a is more polarized than b in the sense of representing less coverage on the spectrum of potential belief.

Property 4: Distinctness refers to the amount the different group distributions can be separated. We can differentiate belief groups by defining meaningful intervals as: ANTI-VAX = [0 - 0.4], NEUTRAL = (0.4, 0.6), PRO-VAX = [0.6, 1]

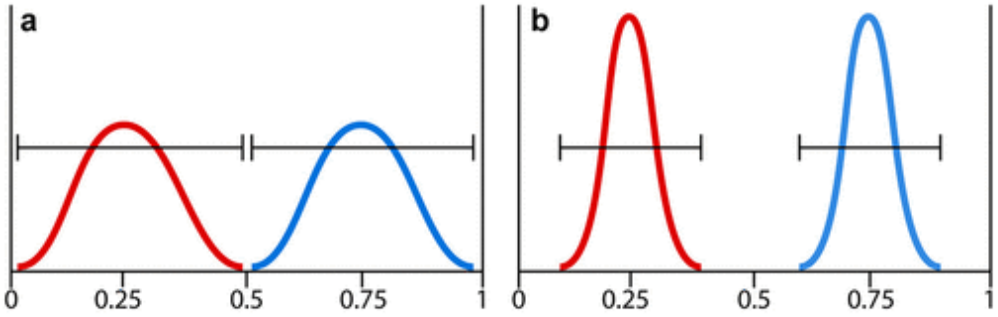
$$Distinctness = Median(PROVAX) - Median(ANTIVAX)$$



Distribution b shows greater polarization than a in terms of distinctness.

Property 5: Group Consensus measures the statistical dispersion in each belief group. Since stances can be spread out anywhere from [0, 1], polarized belief groups tend to be closer to the central idea and therefore each belief group should have a small statistical variance. The smaller this sum for group consensus measurement the greater the polarization.

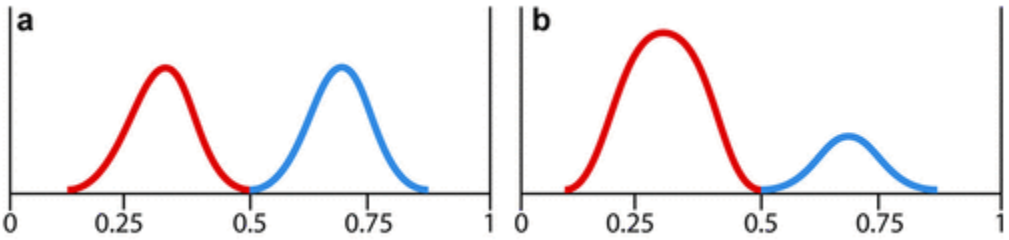
$$Group\ Consensus = \sigma(ANTIVAX) + \sigma(PROVAX)$$



Belief distribution b shows greater polarization than a in the sense of group consensus.

Property 6: Size of groups, polarized distributions often create different belief groups that are equivalent in size.

$$Size = |PROVAX| + |ANTIVAX|$$



Distribution A is more polarized since the two belief groups are roughly the same size.

Finally these different calculations of these properties are combined to create a vector **v**. The magnitude of the vector is calculated, where this magnitude can be used as a single metric to quantify how polarized a topic is. We can compare and rank different topics based on this magnitude. Assuming that all of the individual measurements are of the same statistical significance and contribution, they are assigned a unit weight.

V = [spread, dispersion, coverage, distinctness, group_consensus, size]

P1 =

$$\sqrt{spread^2 + dispersion^2 + coverage^2 + distinctnesss^2 + group^2 + size^2}$$

Alternatively, we can output a qualitative label from a set of discrete categories: Non-Polarized, Moderately Polarizing, Polarized, and Highly Polarized. For each of the six different calculations defined above, we can create threshold values and determine if that property accounts toward polarization. These threshold values were defined from intuition based on several other topics studied. Also these threshold values can be

modified easily. Shown below is an example of how we can output a category for polarization and the rule based conditions for the thresholds.

	Measure 1	Measure 2	Measure 3	Measure 4	Measure 5	Measure 6	total count
topic 1	0	1	1	0	1	1	4
topic 2	1	1	0	1	1	1	5
.....							
topic n	0	0	0	0	0	0	0
0 = does not meet constraint							
1 = meets constraint							
Option 1: discrete categories							
5-6: Highly Polarized							
3-4: Polarized							
2: Moderately Polarized							
0-1: Not Polarized							

```
count_conditions = 0

if(spread >= 0.8):
  count_conditions += 1

if(results_array.std() > 0.33):
  count_conditions += 1

if(distinct_measure >= 0.6):
  count_conditions += 1

if(coverage >= 0.3):
  count_conditions += 1

if(anti_vax_std <= 0.15 and pro_vax_std <= 0.15):
  count_conditions += 1

if(abs(anti_vax_size - pro_vax_size) <= (0.1)*(anti_vax_size + pro_vax_size)):
  count_conditions += 1
```

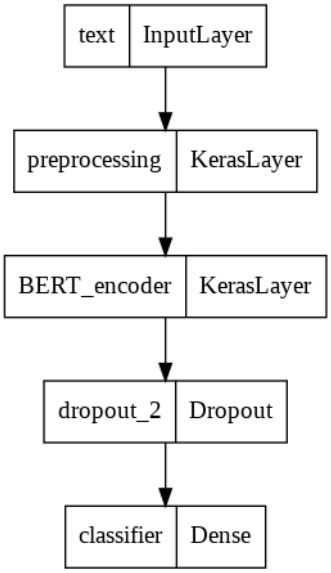
Additionally, we can also take a look at the sentiment analysis of the tweets since polarized and controversial topics tend to have highly emotional views. We can implement sentiment analysis using a well known module, VADER(Valence Aware Dictionary and sEntiment Reasoner) from Python's NLTK(Natural Language Toolkit). I specifically chose VADER for this sentiment analysis as it performs extraordinarily well in social media domains and its dictionary was made and annotated heavily from twitter data by trained and paid Amazon Mechanical Turk workers.[5]

5. Results

Following the methodology detailed above, we analyze the topic of covid vaccination and reveal its polarity. Tweets regarding covid vaccination were gathered from twitter from January 2021 until January 2022. A total of 10,000 tweets were gathered and were then preprocessed. We can label tweets according to the hashtags they contain. For example, if a tweet contains the hashtag, '#vaccineswork' this would be a pro-vax stance and it would get labeled with a 1. After this labeling method, we were left with around 3,000 tweets.

With this labeled dataset, we fine tuned a small BeRT model that contains over 28 million network parameters that are trainable. The BeRT model was fine tuned for the purpose of binary stance classification and trained on the curated covid dataset constructed from the tweets collected earlier. The BeRT model was implemented with Tensorflow and the Keras API; to prevent overfitting, a dropout layer was utilized and used a small learning rate of 3e-5 and 5 training epochs as recommended by the research paper from Google, as well as the AdamW optimizer algorithm[6]. AdamW

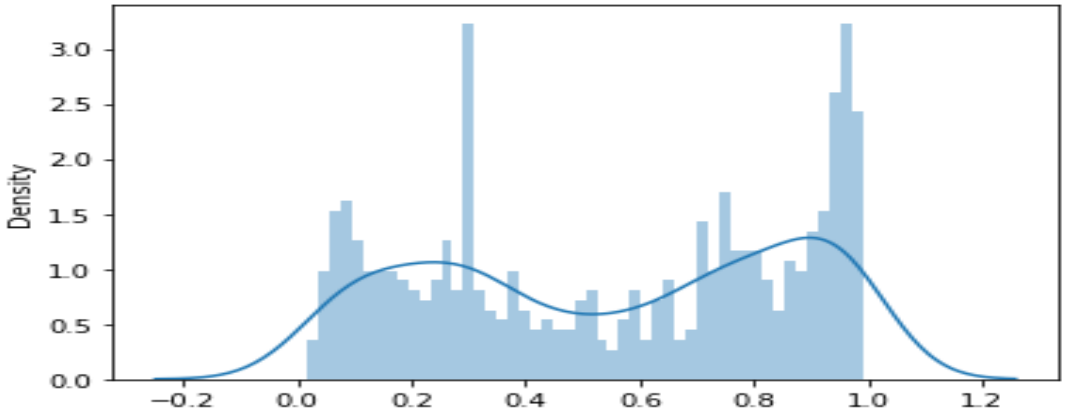
optimizer essentially decouples the weight decay from the optimization step. This means that the weight decay and learning rate can be optimized separately, so changing the learning rate does not change the optimal weight decay. Shown below, the left figure is the fine-tuned BeRT model network architecture, and the figure on the top right demonstrates several custom text stance detection with their respective stance detections performed by our BeRT model. The bottom figure is the distribution of the stances that was predicted by our model with test tweets from the original dataset. The X-AXIS represents the calculated stance/belief. Finally we have our distribution of beliefs to study and analyze and perform the six calculations discussed in methodology.




```
examples = [
    'everyone should get vaccinated they save lives and they work and are safe',
    'vaccines are good!',
    'people who are vaccinated are stupid',
    'vaccines are harmful do not get vaccinated they kill',
    'expose the truth no masks no vaccines',
    'i got vaccinated today and i feel great',
```

```
#1: PROVAX STANCE
#0: ANTIVAX STANCE
```

```
input: everyone should get vaccinated they save lives and they work and are safe : score: 0.899951
input: vaccines are good!            : score: 0.661197
input: people who are vaccinated are stupid : score: 0.478935
input: vaccines are harmful do not get vaccinated they kill : score: 0.486034
input: expose the truth no masks no vaccines : score: 0.214180
input: i got vaccinated today and i feel great : score: 0.823324
```



As shown in the belief distribution for covid vaccination related tweets directly above, we clearly see two segregated and equal sized belief groups; towards the left end of the distribution are the anti-vaccination beliefs, and towards the right end of the distribution are the pro-vax beliefs. We can consider the central valley as having more of a neutral stance towards the topic of vaccination.

Next we calculate the six properties of polarization based on the formulas provided in the methodology section. We first calculate the magnitude of the six dimensional vector, and then we run the rule based conditions on each property to see if it passes a certain threshold. This will give us a categorical representation of how polarized the topic is. Shown in the figure below, are the six individual calculations of each property for polarization, the magnitude of the vector, and the qualitative description all for the topic of vaccination.

```
Below are the calculations for each of the 6 properties defined for polarization
This all relates to the topic of vaccination during the covid-19 pandemic.

The spread of this dataset: 0.972287118434906
The standard deviation of this dataset: 0.3234701156616211
The coverage of this distribution is: 0, there are no empty bins
The measure of distinctness between the pro vaccination beliefs and anti vaccination beliefs: 0.6898642927408218
Anti_Vax standard deviation(dispersion): 0.10788021241910156, Pro_Vaccination standard deviation 0.11021197085351325
Size of anti vax belief group: 243, Size of pro vax belief group: 267

The Magnitude of the 6D Vector: 510.0015425977643
Polarized
4/6 of the conditions were met for this belief distribution
```

Finally, we can run the sentiment analysis algorithm, VADER, to check the assumption that polarized topics tend to be emotionally fueled and have strong sentiments. The VADER algorithm was run on the covid tweets and the X-Axis represents the sentiment analysis score, where -1 is a negative sentiment, and +1 is a positive sentiment. We can see the distribution strongly resembles the belief distribution that was constructed from the BeRT stance detection classifier as well as the labeling method discussed in methodology.

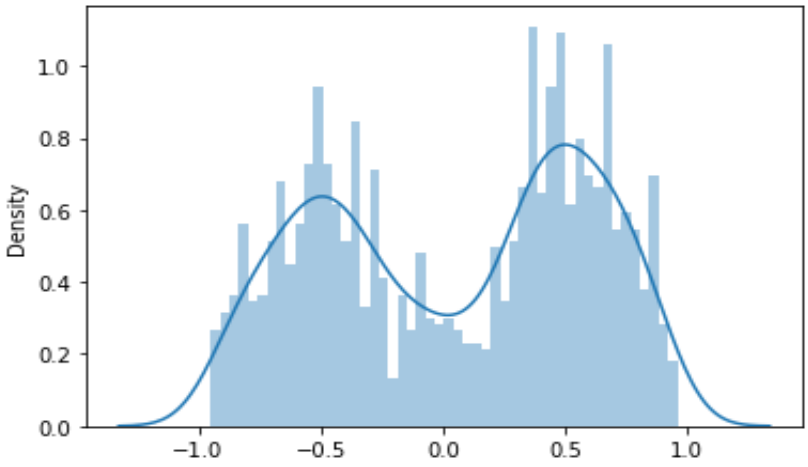


Table of results for tested topics:

	Covid Vaccination	**Gun Rights**	**Abortion**
# of Tweets	3,000	5,800	5,690
Polarization Index	510	1152	975
Category	Polarized	Highly Polarized	Highly Polarized

Based on the polarization index(magnitude of the 6D vector), the largest index will be the most polarized topic, therefore, 1.) Gun Rights 2.) Abortion 3.) Covid Vaccination

6. Conclusion

In conclusion, the results of the experiment from the methodology, are specifically good for known polarized topics such as: vaccination, abortion, gun rights. I implemented the methodology on the topics listed above and achieved fairly similar results as that of the vaccination topic that was discussed throughout this paper. The only issue with the latter two topics is that tweets were collected in a short time window as opposed to an entire year of tweets for covid.This short window for collection of tweets meant that similar language was being picked up and trained for BeRT causing it to overfit. Visually outputting the belief distributions and sentiment analysis gives us a quick and clear picture of the topic. A major limitation is that this procedure is not compatible with known non-polarized topics. However, with this procedure in this paper, we can perform custom stance detection, quantify a polarization index, and even rank different topics based on these calculations. Motivations for social polarization on the web is that it is a dominant aspect of conflict such as hate speech and misinformation spread. With better improvements in computational methods for social polarization, we can measure, understand, monitor, identify and even mitigate misinformation spread on social media platforms. It can help to identify future political campaigns and prevent the spread of misinformation. As well as improving the classification performance on hate-speech. Social polarization is still a topic of research and methods for extracting, detecting, and quantifying it are an area of research.

7. Future Works

After completion of this project, we now discuss some ideas for future work to be done. Initially I used the small fine-tuned BeRT model for binary stance detection having a sigmoid activation function in the last output layer. For example if we want to have a multiclass classification for a topic with more than two beliefs, we can simply alter the activation function to be softmax to output a vector of probabilities. Potentially using the larger BeRT model with over 100 million trainable parameters will improve stance detection. For this project, the threshold values were intuitively chosen based on three different experiments I ran with various topics and I analyzed the distributions and chose meaningful threshold values. For future work, we need to figure out the actual cut-off threshold values based on evidence from several experimental runs or a way to fine tune them. Also the 6D Vector of the properties of polarization all have an equivalent unit weight, therefore each of them have the same significance and contribution to the final output of the magnitude of the vector. Some of these six different properties may have differing weights as one is more important than the other. The size of the belief groups has the majority of the contribution to the final output as it is the only measurement to be greater than 1.This too is something that needs to be figured out for future works. Finally, running this experiment and analysis takes a considerable amount of time to gather the tweets, process them, fine tune a BeRT model and analyze the belief distributions. To circumvent this, we can deploy this project on the cloud, and continuously run it in the background.

8. References

[1]. Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Flocken, C., & Holman, B. (2022, January 1). *Understanding polarization: Meanings, measures, and model evaluation: Philosophy of science*. Cambridge Core. Retrieved July 20, 2022, Retrived from https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/understanding-polarization-meanings-measures-and-model-evaluation/FFACA58EFC4A6E421107020290BF592C#_i2

[2]. Brian Sharber. (December 15, 2020). Analyzing Political Polarization in News Media With Natural Language Processing. Retrieved from https://jewlscholar.mtsu.edu/server/api/core/bitstreams/f6cae15a-60d8-45d2-b3ce-c287fcdb984f/content

[3]. Pedro H. Calais Guerra, Wagner Meira Jr, Claire Cardie, and Robert Kleinberg (2013, March 21). A Measure of Polarization on Social Media Networks Based on Community Boundaries. Retrieved from https://www.cs.cornell.edu/home/cardie/papers/ICWSM13-Polarization.pdf

[4]. Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, Michael Mathioudakis (2017, September 20). Quantifying Controversy in Social Media. Retrieved from https://arxiv.org/pdf/1507.05224.pdf

[5]. Hutto, C. J., & Gilbert, E. (2014, May 16). *Vader: A parsimonious rule-based model for sentiment ... - Eric Gilbert*. Eric Gilbert Papers. Retrieved from http://eegilbert.org/papers/icwsm14.vader.hutto.pdf

[6]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Retrieved from https://arxiv.org/pdf/1810.04805.pdf