Quantifying Changes in Social Polarization Over Time and Region

David L. Edwards

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Scotland Leman, Chair

Jennifer Van Mullekom

Jyotishka Datta

James E. Hawdon

July 25, 2024

Blacksburg, Virginia

Keywords: Topic Models, Polarization, News, Sentiment Analysis, Time Series

Quantifying Changes in Social Polarization Over Time and Region

David L. Edwards

(ABSTRACT)

Recent studies indicate that Americans have grown increasingly divided and polarized in recent years [BGS22], [Haw+20]. This research aims to describe and measure polarization trends across a historical archive of US-based, primarily regional, newspapers. The newspapers chosen are from various US markets to capture any regional differences in the discussion of issues/topics. Our modeling approach employs the Structural Topic Model (STM) to identify topics within a given corpus and measure the tonal differences of articles discussing the same topic. Specifically, we use the STM to infer potentially related articles and a sentiment analyzer called VADER to identify topics with a high level of semantic disparity. Using this method, we assess the polarization of developing and evolving topics, such as sports, politics, and entertainment, and compare how polarization between and within these topics has changed over time. Through this, we create topic-specific sentiment distributions, referred to as polarization distributions. We conclude by demonstrating the usefulness of these distributions in identifying polarization and showing how high polarization aligns with significant social events.

Quantifying Changes in Social Polarization Over Time and Region

David L. Edwards

(GENERAL AUDIENCE ABSTRACT)

Most Americans have a sense that their nation is becoming more socially polarized. Numerous studies and anecdotal evidence supports this. Our aim with this work is develop a method to quantify polarization in text media and apply this method to news articles published in local and national newspapers. Using a statistical model we are able to group articles based on a common shared topic. We then analyze the sentiment of each article and evaluate how sentiments for a particular topic change over time. We then compare newspapers based on location, political endorsements, and ownership groups.

Dedication

This work is dedicated to my sons, Lucas and Zach. You inspire me everyday and I'm so lucky to be your father.

Acknowledgments

This work would not be possible if not for the support and understanding of my partner, Jennifer Johnson along with my parents and children. I would also like to think Dr. James Hawdon for his help in verifying my early results and helping me understand the sociology theory which inspired this work. Finally, none of this would be possible without the support and dedication of my advisor, Dr. Scotland Leman. No other advisor could have gotten me through this work and helped me navigate my personal challenges. Thank you all!

Contents

List of Figures

List of Tables

List of Abbreviations

AMT Amazon Mechanical Turk

DT Decision Trees

EM Expectation Maximization

LDA Latent Dirichlet Allocation

LSI Latent Semantic Index

MCMC Markov Chain Monte Carlo

ML Machine Learning

NB Naïve Bayes

NLP Natural Language Processing

pLSI Probabilistic Latent Semantic Index

STM Structural Topic Model

SVD Signular Value Decomposition

SVM Support Vector Machine

VADER Valence Aware Dictionary for sEntiment Reasoning

NLP is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages.

LSI is a model developed by Deerwester et. al. [Dee+90] that uses Singular-Value-Decomposition (SVD) to characterize and group documents with in a corpus.

SVD is a method to decompose a matrix (non-uniquely) into components that consist of orthonormal vectors and a diagonal matrix. It is akin to eigen decomposition of a square matrix but is applicable to any size matrix.

pLSI is a model developed by Hofmann [Hof99] that extends LSI and using a probabilistic frame work to group documents.

EM is a method of maximizing a likelihood function via and iterative process.

LDA is the first model to explicitly define the latent space as topics.

MCMC is a collection of algorithms designed to obtain samples from probability distributions.

STM is a topic model that builds upon LDA by including covariates and correlated topics.

ML is a broad category of classification algorithms.

NB is an ML algorithm that uses Bayes rule to classify objects.

SVM is an ML algorithm that classifies objects based on hyperplanes which separate observations.

DT is an ML algorithm that uses trees to classify objects.

VADER is an unsupervised sentiment analyzer.

Chapter 1

Introduction

By nearly all accounts, whether anecdotal or scientific, Americans are becoming more socially polarized. See [Cam18], [Kle20], [JJM18], and [Han11] for a sampling of the many books and articles written on this subject. Much research has gone into understanding the causes and effects of this hyper-polarized environment. However, this work focuses on developing a method to quantify the polarization and using this quantification to detect moments of high and low polarization. We take a topic-centric approach, which allows us to focus on specific polarizing events and times based on how they affect a given topic. We combine the results of a topic model with a sentiment analysis that quantifies the underlying tone of a text. In addition, we developed an automated topic-matching procedure, which allows us to connect topics over time. The combination of these tools, topic modeling, sentiment analysis, and topic matching, allow us to detect tonal shifts in the discourse surrounding topics, and we show that these tonal shifts correspond to periods of high or low social polarization.

1.1 Disertation Outline

This dissertation comprises background, definitions, methodology, and results based on real-world data. In Chapter 2, we present the background necessary to understand topic modeling and sentiment analysis. We present the various evolutions and advances in topic modeling and justify our selection of the structural topic model for our analysis. Finally, we end the chapter with a review of current techniques used in sentiment analysis and explain the necessity of using an unsupervised sentiment analysis tool.

In Chapter 3, we analogize the polarization in the news media to the polarization of light and use this analogy to help inform the signatures of social polarization we can investigate. Then, we define polarized and unpolarized topics based on the sentiment used to discuss the said topic. In Section 3.2, we introduce the sociological theory that motivates much of this work and how we use the concept of social capital to hypothesize moments of low and high polarization. Finally, we conclude the chapter by reviewing the ethics associated with statistical practice and discussing the steps to ensure ethical analysis and conclusions.

In Chapter 4, we describe our methodology in detail, discussing the data source and selection, how we combine sentiment analysis and topic modeling to generate the polarization distribution, and the post-processing required to add human interpretation to the topics and match the topics over time. This chapter presents our method in general terms to show that our method is translatable to other corpora or domains.

Chapter 5 presents the results of applying our method to a data set of newspaper articles obtained from the NewsBank database. We investigate what impact certain hyperparameters have on the potential conclusions drawn from the analysis. Then, we generate time series plots for both anticipated high- and low-polarization moments and investigate interesting changes in these plots, looking for precipitating events. Finally, we end the chapter by investigating possible differences between newspapers based on region, political endorsements, and ownership groups.

Chapter 2

Review of Literature

2.1 Topic Modeling

In 1990, Deerwester et al. [Dee+90] published an article on improving document retrieval methods based on keyword searches. They introduced Latent Semantic Indexing (LSI), which matches documents using spectral decomposition of the term-document matrix. Their goal was not to model or extract topics but to enhance document matching beyond traditional keyword methods. Keyword-matching involves selecting keywords to represent documents, which Deerwester et al. found inefficient due to words having different meanings in different contexts. They aimed to differentiate these instances. Although they did not claim to model topics, they used the concept of topics to guide their work, as indicated by: "We might consider any document (or title or abstract) to consist of a small selection from the complete discourse that exists for a given topic. Thus, the text from which we extract index terms is a fallible observation from which to infer what terms apply to its topic." The word "topic" appears only three times in the article, twice in the above quote. Instead of "topic," they used "factor(s)," a term common in Singular Value

Decomposition (SVD), the method they used.

Probabilistic Latent Semantic Indexing (pLSI) [Hof99], an evolution of LSI, places LSI's "factors" on a probability scale, adding interpretability. Like LSI, pLSI aims to improve automated document indexing, using "topics" as a guiding principle. Modeling topics probabilistically enables the use of a likelihood function and statistical methods like the Expectation Maximization (EM) algorithm to fit the model.

Topic modeling gained popularity with Blei et al.'s paper *Latent Dirichlet Allocation* (LDA) [BNJ03]. LDA builds on Hofmann's pLSI by adding a Dirichlet prior over document topic proportions, placing it in a Bayesian context. This makes LDA a generative model capable of computing the probability of a document given a corpus. The conjugate prior over topic/document proportions allows for various estimation methods, such as methods that require the use of Markov Chain Monte Carlo (MCMC) and Gibbs sampling algorithms. Blei et al. emphasize modeling topics and provide an interpretation of their probability distributions over the vocabulary. This seminal paper spurred numerous LDA modifications, including correlated topics and time-based models. See [BL06a], [BL06b], [LM06], [Qua+15], and [WM06] for examples.

One such topic model developed in the years following the publication of LDA is the Structural Topic Model (STM) by Roberts et al. [RSA16]. STM uses the same base topic/word model as LDA but extends this model to allow for document-level and topic-level covariates and the ability to model the correlation between topics to quantify how frequently they appear in the same document. STM was developed

primarily with social scientists in mind. The addition of covariates to the model allows scientists to model various features and chart their relationship to the topics.

2.1.1 LSI

Deerester et al. developed LSI to solve a specific problem: improving document retrieval methods based on user-specified words and phrases. At the time and still to this day, it was common to represent a corpus as a matrix of T terms by D documents. The entries i, j in this matrix are integer values representing the number of occurrences of the term t_i in the document d_j. Before LSI, standard document retrieval methods included pulling documents that matched the highest number of terms in a search phrase, linking documents with keywords, and matching documents to a search phrase based on these keywords. Deerwester et al. illuminate two problems associated with these methods. The first is what the authors label as the problem of *synonymy*, namely that there are several ways to refer to the same object, concept or expression. The words in the document (or keywords) and search words might differ, while the searcher intended to use them to represent the same topic or idea. This notion of miss-specified terms is related to type II error or false negative; the ground truth is that the document and the search should have matched, but the algorithm did not match them. The second problem is what they call the *polysemy*, which refers to the fact that most words have different meanings depending on the context in which they appear. It is possible for the document's authors and the search's authors to use the exact words but have different desired topics or concepts

in mind. This notion is related to a type I error or a false positive. The truth is that the document does not match the search, but is labeled as if it were a match.

LSI attempts to solve both of these problems by removing what the authors refer to as "noise" and attempting to analyze the underlying message conveyed in a document without focusing on the precise words chosen by the author(s) of the document. LSI attempts to solve the problems of synonymy and polysemy in the same way by "matching" documents based on the terms used within each document and the latent set of terms associated with the collective context of the document. In doing so, LSI solves the *synonymy* problem by allowing interchangeable words to occupy the same latent space. Furthermore, LSI solves the *polysemy* problem by allowing single terms to occupy multiple latent spaces depending on context.

Deerwester et al. use the term-document matrix's Singular Value Decomposition (SVD) to model this latent space and remove the "noise" associated with the term-document matrix. SVD is a generalization of the eigen-decomposition of a square matrix. There are a few nearly equivalent ways to define SVD. As with almost anything in mathematics, if one places enough constraints on the said object, such as the decomposition of the term-document matrix, one can obtain a unique decomposition. For our purposes, we will define the SVD in terms of the document-term matrix. Suppose that we have a document term matrix, X, of size $M \times V$ where M is the number of documents and V is the number of terms. Note that X is typically an extremely sparse matrix, as within any corpus of moderate size, the overwhelming majority of vocabulary terms will not appear in any given document. The singular value decomposition of X is the product of three matrices:

$$X = DST^t, \tag{2.1}$$

where D is a $M \times R$ matrix where the columns are orthonormal, T is a $V \times R$ matrix again with orthonormal columns, and S is an $R \times R$ diagonal matrix. Here, we assign R to the rank of X. As mentioned, there are alternative ways to define the three matrices in the decomposition. However, defining the matrices in this way and adding the constraint that the elements in S appear along the diagonal in decreasing order are enough to ensure the uniqueness of the decomposition up to a sign change. That is $\forall i, j \in \mathbf{Z}$ such that $1 \leq i < j \leq r$, we have $S_{i,i} \leq S_{j,j}$.

The LSI authors used this decomposition to model the latent space of a given corpus. They propose only keeping the K largest values in the diagonal matrix S. The intuitive justification for this move is that the remaining terms are the "noise" the authors want to remove. However, the authors do not give any way to estimate K. As we shall see, this will be a recurring theme for topic models. Most topic modeling relies on a user-defined number of topics, with no way to estimate this value from the data. As is often the case with hyper-parameters, cross-validation techniques are one option that many use to select the number of topics that minimize a given metric for out-of-sample documents. The authors of LSI make a recommendation based on their experience setting $K = 50$ or 100. In this way, the three matrices that make up the singular value decomposition of X are truncated so that D is approximated by $\hat{D}$ a matrix $M \times K$, T is approximated by $\hat{T}$ a matrix $v \times K$ and S is approximated by $\hat{S}$ a diagonal matrix. $K \times K$. Thus, the estimate for the full document-term matrix

is given by:

$$X \approx \hat{X} = \hat{D}\hat{S}\hat{T}^t. \tag{2.2}$$

Deerwester et al. point out that $\hat{X}$ is the K rank model with the best fit of least squares to X. In this way, the representation of all the documents and terms in the corpus is points in a K dimensional space, where the matrix $\hat{D}$ gives the dimensions for each document, and the matrix $\hat{T}$ gives the dimensions of each term. With this framework in mind, we can understand the proposed solution of the author to the problems of *synonymy* and *polysemy*. If a term is polysemic, it will map onto multiple dimensions in this K space where the different dimensions or factors represent the different meanings of the term based on context. Furthermore, the authors use this representation to solve the problem of *synonymy* as terms that are interchangeable in a given context will be mapped onto those same factors.

It is important to note again that the authors of the LSI model never intended to model topics or perform any natural language processing. Their main goal and focus was to improve the recall and precision of document retrieval methods. The authors skirt around the notion of an underlying latent topic space for a given corpus, but directly push against any desire to interpret or understand the underlying factors. Saying, "Our aim is not to be able to describe factors verbally but rather to be able to represent terms, documents, and queries in a way that escapes the unreliability, ambiguity, and redundancy of individual terms as descriptors" [Dee+90].

2.1.2 pLSI

The next notable improvement in topic modeling came with the introduction of the Probabilistic Latent Semantic Indexing (pLSI) model by Hofmann [Hof99]. As the name implies, just as with LSI, pLSI attempts to define a latent space by which to represent both words and documents in a given corpus. However, instead of using SVD to consider the spectral decomposition of the document-term matrix to estimate this latent space, pLSI proposes a mixture model that clusters the set of words and documents. As with LSI, one of the goals of this latent space is to make associations between words, *synonymy*, and to distinguish when a word has different meanings depending on the context in which it is used, *polysemy*. In comparison, LSI solved these problems by projecting the observed term-document matrix onto the most significant K components and using these components to make associations between words, pLSI clusters words, and documents into what the author describes as "factors," which correlate both words and documents. In this way, pLSI mitigates the issue of *synonymy* by including the latent classes, which cluster words based on cooccurrence in the same documents. Thus, if two or more words are interchangeable, the words will co-occur with a similar set of words and thus be assigned a similar value to a given cluster. Additionally, it solves the *polysemy* problem by allowing words to appear in all clusters, which allows the other words in that cluster to define the use of that word based on context.

The pLSI model is relatively straightforward to describe. Given a corpus of M documents labeled $d_i, i = 1, \ldots, M$ and a vocabulary of V words labeled $w_j, j =$

$1, \ldots, V$, there exist K latent class variables labeled $z_k, k = 1, \ldots, K$. The joint probability model is presented in equation 2.3.

$$\begin{aligned} P(d_i, w_j) &= P(d_i)P(w_j|d_i), \\ P(w_j|d_i) &= \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i). \end{aligned} \tag{2.3}$$

This is equivalent to the following generative model:

1. Sample a Multinomial distribution to select a document d based according to probabilities $P(d)$,

2. For each word in d sample a Multinomial distribution to select the latent variable z according to probabilities $P(z|d)$,

3. Sample a Multinomial distribution with probabilities $P(w|z)$ to generate a word.

From equation 2.3, we see that this model is a mixture model of multinomial distributions. In this model are implied two independence assumptions. First, words from the same document are assumed to be conditionally independent of each other. That is, given document d_i and words w_j and w_h within d_i we have $P(w_j, w_h|d_i) = P(w_j|d_i)P(w_h|d_i)$. This assumption is often referred to as the "bag of words" assumption, as it implies that under this model, documents are a multiset of words in which the order of the words is unimportant. The second assumption is that,

conditioned on the latent variable z, the generation of words is independent of the document d. Notationally, this is equivalent to $P(w|z,d) = P(w|z)$.

This model has a geometric interpretation that adds intuition and helps illustrate the data reduction such a model provides. As mentioned above, pLSI is a multinomial distribution mixture model. The K clustering distributions, $P(\cdot|z_k)$, are probability vectors of length V and, as such, must sum to 1 for a given cluster, and each value within the vector must be positive. That is $\sum_{j=1}^{V} P(w_j|z_k) = 1$ and $P(w_j|z_k) \geq 0$ for all w_j, $j = 1, \ldots, V$ and for all z_k, $k = 1, \ldots, K$. Given this constraint, the possible values for the vector, $P(\cdot|z)$, are limited to the $V-1$ simplex. A simplex is a generalization of a triangle to dimensions other than 2D. For simplicity's sake, suppose that the corpus under consideration consisted of only three vocabulary words. In this rather trivial example, $P(\cdot|z)$ is a 3-vector that can take on any value within the 2D simplex shown in the red triangle in Figure 2.1. This triangle has vertices at $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$. As mentioned above, this idea can be extended to higher dimensions. For example, suppose that our corpus consists of four words. The values that $P(\cdot|z)$ can take are represented by the 3D surface in the 4D space that connects the four vertices $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$, and $(0,0,0,1)$. Thus, for a corpus of size V, the $V-1$ simplex fully describes the values that $P(\cdot|z)$ can obtain. These statements are also valid for $P(\cdot|d)$.

Assuming that we have a corpus of M documents with a vocabulary of size V, the vectors $P(\cdot|d)$, along with the size of each document, thoroughly describe the data given the two assumptions mentioned above. Data reduction comes from finding the K vectors, $P(\cdot|z)$, in the $V-1$ simplex whose span is closest to the M vectors of

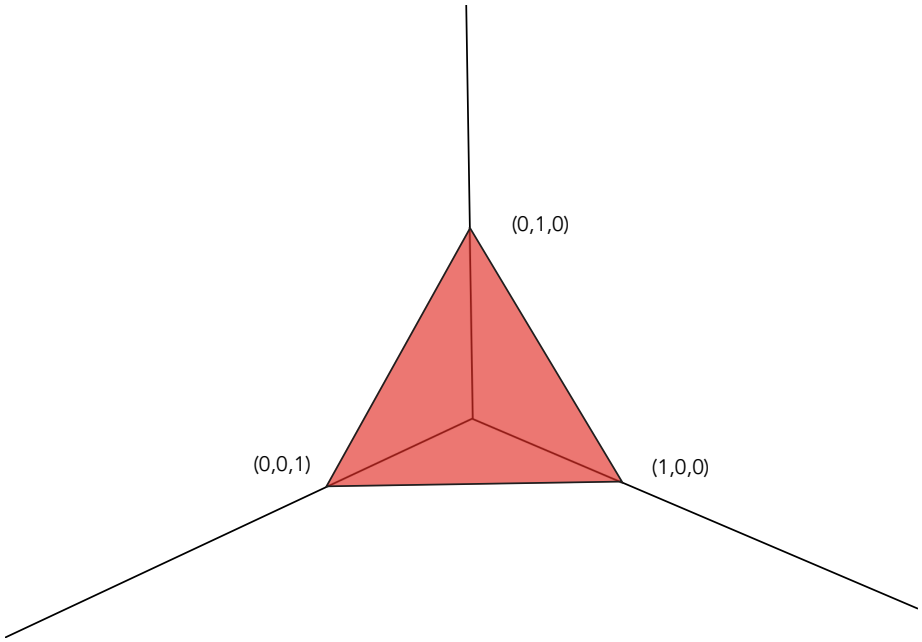


Figure 2.1: 2D Simplex embedded in 3D space.

$P(\cdot|d)$. Assuming that $K \ll M$, this can be several orders of magnitude reduction in the data size.

As a simplified example that is easy to visualize, consider a corpus of ten documents that consist of three words. The limited number of words is necessary to plot the vectors $P(\cdot|d)$ in a 3D space. Figure 2.2 shows a scatter plot of these ten vectors. Notice that all the points are in the two simplex shown in Figure 2.1. If we want a set of vectors that span the given collection, we could use $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$, but this does not allow for any data reduction and results in each of the topics representing a single word. Thus, we will model these ten documents using two latent topics represented by vectors whose span is the "closest" to these documents. Given that this model requires simultaneously estimating the vectors that form the

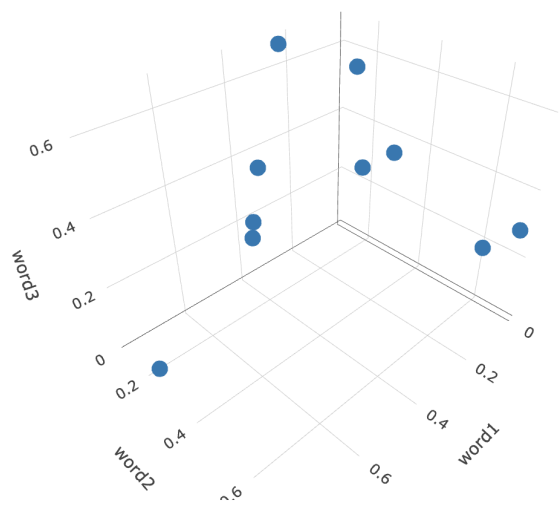


Figure 2.2: 10 Points in the 2 simplex that represent word/document proportions.

topics and the document proportions, the Expectation Maximization (EM) algorithm works well in estimating these unknowns.

The EM algorithms alternate between two steps. These steps are labeled the E-step for Expectation and the M-step for Maximization. To fit the pLSI model, the E-step involves estimating the latent topic classification for each word/document combination, that is:

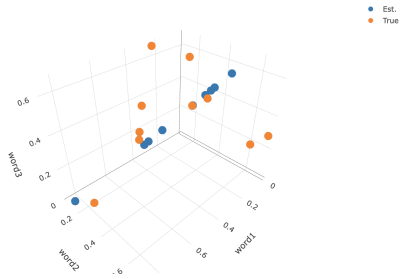
$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{\tilde{z}} P(\tilde{z})P(d|\tilde{z})P(w|\tilde{z})}. \tag{2.4}$$

The M-step involves reestimating the spanning vectors (topics) and the coefficients for the vectors (topic/document proportions) based on the updated classification

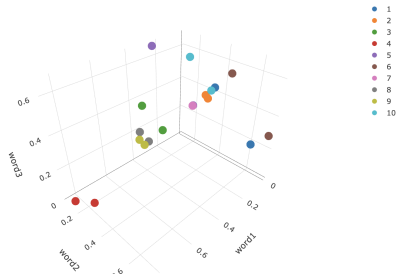
topic probabilities obtained in the E-step. Thus, the EM algorithm proceeds by alternating between equations 2.4 and 2.5. The algorithm concludes when the quantities of interest, $P(w|z)$ and $P(d|z)$, do not change from one iteration to the next.

$$
\begin{aligned}
P(w|z) &= \frac{\sum_d n(d,w)P(z|d,w)}{\sum_{\tilde{w}} \sum_d n(d,\tilde{w})P(z|d,\tilde{w})} \\
P(d|z) &= \frac{\sum_w n(d,w)P(z|d,w)}{\sum_{\tilde{d}} \sum_w n(\tilde{d},w)P(z|\tilde{d},w)} \\
P(z) &= \frac{\sum_{d,w} n(d,w)P(z|d,w)}{\sum_{d,w} n(d,w)}
\end{aligned}
\tag{2.5}
$$

Applying the EM algorithm to the ten hypothetical documents presented in figure 2.2 assuming two "topics" gives the estimated word/document proportions in figure 2.3a. One exciting feature of these estimates is that they are all co-linear. These estimates are linear combinations of two vectors in the two simplex, and the span of any two vectors is a plane through the origin and the endpoints of the vectors. However, we have the added constraint that all these estimates must lie in two simplexes. Thus, all estimates lie on the intersection between the plane spanning the topic vectors and the required simplex, which is why all the estimates are colinear in this example. The data values and their estimates are shown again in figure 2.3b, where each pair of actual data values and their estimates are color coded to show how well the model estimates the data.



(a) Actual data points are in orange. The pLSI model estimated points are in blue.



(b) Each color pairs up the actual data point with the pLSI model estimated point.

2.1.3 Latent Dirichlet Allocation

Topic models gained popularity after the publication of the Latent Dirichlet Allocation (LDA) model by Blei et al. [BNJ03]. Although this model is not dramatically different from the pLSI model presented by Hofmann, the authors are among the first to understand that the latent space created in the pLSI model could be interpreted as "topics" within the given corpus. Also, along these same lines, it appears that Blei et al. are the first to define topics as a probability distribution over words. In any event, Blei et al. saw the utility of the pLSI model beyond improving document retrieval. They popularized the LDA model as a powerful tool for anyone analyzing an extensive collection of documents.

The LDA model is given below as a generative model and a plate diagram in Figure 2.4.

1. Choose $N \sim \text{Poisson}(\xi)$

2. Choose $\theta \sim \text{Dirichlet}(\alpha)$

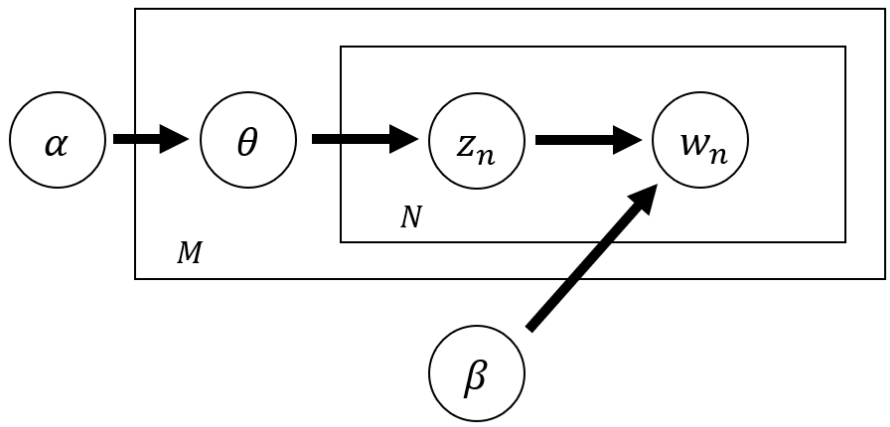


Figure 2.4: Plate Diagram for the LDA model. The rectangles represent replicates, the outer rectangle is for the M documents and the inner rectangle is for the N words within each document.

3. For each of the N words in the document:

   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$

   (b) Choose a word $w_n \sim \text{Multinomial}(\beta_{z_n})$

The observant reader will notice that the only difference between the LDA generative model and the pLSI generative model is that LDA explicitly adds the Dirichlet distribution to describe the generation of the topic/document proportions, as the model name would suggest. The Dirichlet distribution is a generalization of the beta distribution to allow for more than two categories of interest. Just as the beta distribution is conjugate to the binomial distribution, the Dirichlet is conjugate to the multinomial distribution. This conjugacy makes Dirichlet the clear choice for the prior on the probabilities of each category of a multinomial, as is the case with topic/document proportions. The parameters of a Dirichlet distribution are

an integrer $K \geq 2$ and a vector $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K$. The Dirichlet distribution probability density function (pdf) is given in Equation 2.6.

$$f(x_1, x_2, \ldots, x_k | \alpha_1, \alpha_2, \ldots, \alpha_k) = \frac{1}{B(\vec{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \tag{2.6}$$

Here, $B(\vec{\alpha})$ is the multivariate beta function. Note that the support for this distribution is the $K-1$ simplex. That is, the set of all vectors $\vec{x}$ such that $x_i \geq 0 \; \forall i$ and $\sum_{i=1}^{K} x_i = 1$. Because of the nature of the support described above, the Dirichlet distribution is called a distribution over distributions. Examples of the Dirichlet distribution with three categories and different values for the concentration parameter $\vec{\alpha}$ are shown in figure 2.5.

Using the Dirichlet as a prior on the topic/document proportions provides modeling benefits such as regularizing the topic/document space, which we shall see an example of below. It allows for data reduction in the document space, whereas the pLSI model uses an empirical estimate of each document's topic proportion. Having such a prior allows for out-of-sample predictions and goodness-of-fit evaluations, such as computing the perplexity of out-of-sample documents. The regularization of the topic/document space is illustrated in figure 2.6, where, once again, we are using a low-dimensional example to help guide intuition. Just as in Figure 2.2, we are again considering a scenario where the entire vocabulary consists of 3 words, and thus all word/document proportions lie on the 2D simplex represented by the blue triangle. As mentioned in section 2.1.2, using three topics in the setting would be pointless as this would result in each word representing a topic and thus would not allow for

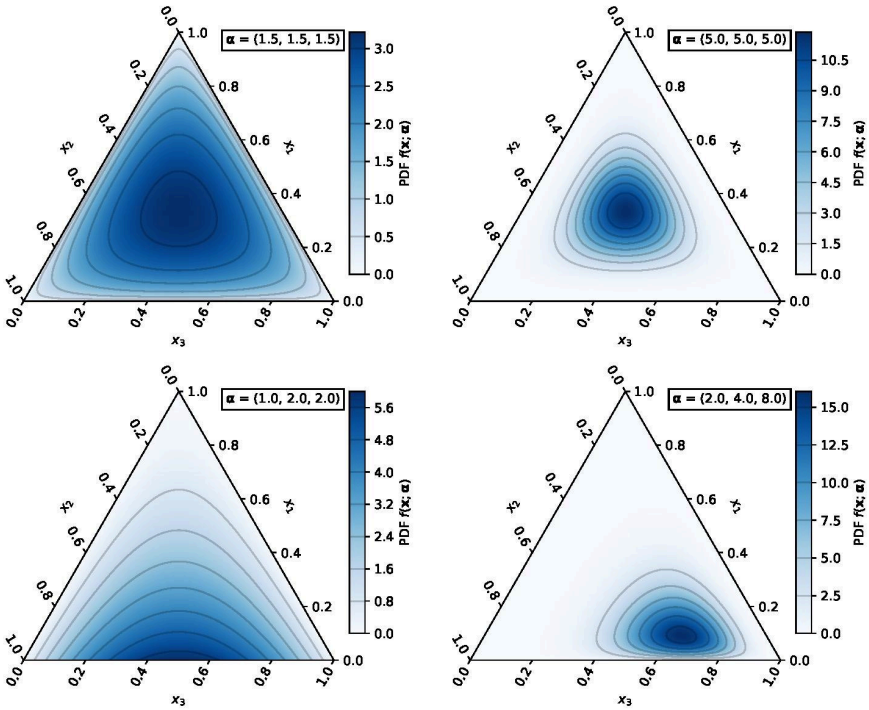


Figure 2.5: Examples of Dirichlet distributions with three categories and different values for the parameter $\vec{\alpha}$

[Ner]

any data reduction. However, for illustration purposes, assume that there are three topics, each represented by the three corners of the red triangle. Each of these points represents the word/topic proportions for each topic. As mentioned in the generative models for both pLSI and LDA, the representation of each document in the corpus is a proportion of each topic. If we multiply the topic proportion of each document by the word proportion of each topic, we get an estimated word/document proportion. Thus, we can plot each document on the red topic simplex shown in figure 2.6. As mentioned, the main difference between the pLSI and LDA models is the Dirichlet prior distribution over the topic/document proportions. The benefit of this prior is easy to understand in the context of Figure 2.6. Under the pLSI model, each document has its empirical topic proportion, but nothing connects or relates the topic proportions across documents. In figure 2.6, these empirical distributions are represented by the green stars within the topic simplex. On the other hand, in LDA, with the addition of the Dirichlet prior, it is assumed that all topic/document proportions are drawn from this prior distribution, thus allowing for smoothing in the document space as represented by the gray ovals. This smoothing is exactly what we would expect given the graphs of the Dirichlet distribution in Figure 2.5.

2.1.4 Structural Topic Model

As discussed at the beginning of Section 2.1, the publication of the LDA model was a turning point for topic modeling. Blei et al. added the interpretation of topics missing from the LSI and pLSI models, which focused on efficient document retrieval

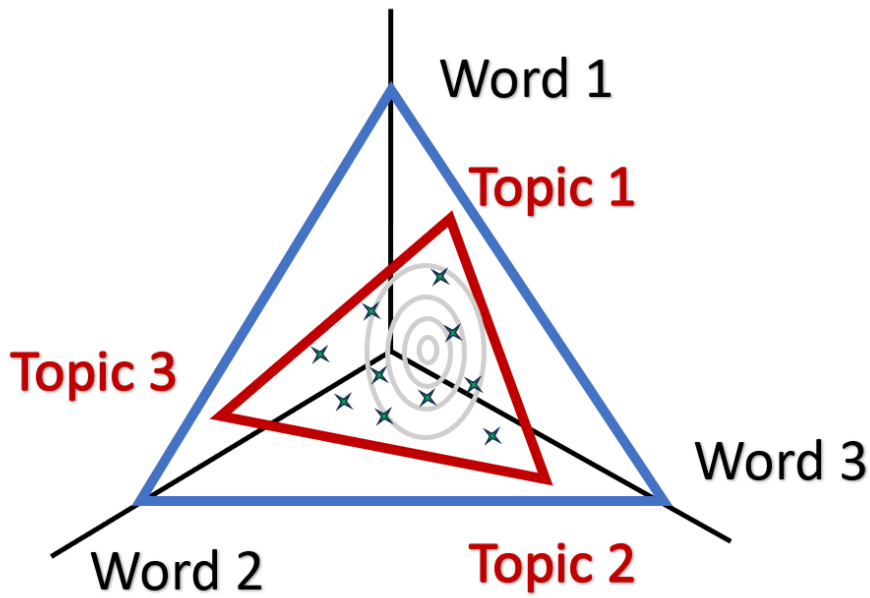


Figure 2.6: The blue triangle represents the three-word simplex embedded in 3D space. The red triangle represents the topic simplex of three topics. Each corner of the red triangle represents a different topic. A topic is defined to be a probability distribution over words. The points within the red triangle represent possible document/topic proportions as mixtures of the three topics that make up the red triangle. The pLSI model treats each document in the corpus as a single point within the topic simplex of possible proportions, represented by the green stars in the figure. The LDA model places a distribution over these possible document topic proportions, represented by the gray ovals.

based on search words. Once this framework and interpretation of topic modeling spread, other authors started working on augmentations and adapting the model to quantify interesting relationships between topics of a given corpus. One of the first enhancements to LDA was to introduce a way to capture possible correlations between topics, as seen in [BL06a] and [LM06]. Another area of interest for modelers was the ability to include covariates in modeling processes. The Structural Topic Model (STM) developed by Roberts et al. incorporates both of these characteristics [RSA16]. The generative model for a document d given K topics according to the STM is shown in Equation 2.11.

$$\gamma_k \sim \text{Normal}_p(0, \sigma_k^2 I_p), \qquad \text{for } k = 1 \dots K-1, \tag{2.7}$$

$$\boldsymbol{\theta}_d \sim \text{LogisticNormal}_{K-1}(\boldsymbol{\Gamma}' \boldsymbol{x}_d', \boldsymbol{\Sigma}), \tag{2.8}$$

$$\mathbf{z}_{d,n} \sim \text{Multinomial}_K(\boldsymbol{\theta}_d), \qquad \text{for } n = 1 \dots N_d, \tag{2.9}$$

$$\mathbf{w}_{d,n} \sim \text{Multinomial}_V(\mathbf{B}\mathbf{z}_{d,n}), \qquad \text{for } n = 1 \dots N_d, \tag{2.10}$$

$$\beta_{d,k,v} = \frac{exp(m_v + \kappa_{k,v}^{(t)} + \kappa_{y_d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)})}{\sum_v exp(m_v + \kappa_{k,v}^{(t)} + \kappa_{y_d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)})} \qquad \text{for } v = 1 \dots V \text{ and } k = 1 \dots K, \tag{2.11}$$

In this model $\mathbf{X} = [\mathbf{x_1}|\mathbf{x_2}| \dots |\mathbf{x_D}]'$ is the $D \times P$ matrix of document-level covariates. These could be continuous variables such as the publication date or categorical variables such as author. The matrix $\boldsymbol{\Gamma}$ has dimension $P \times K - 1$ and represents the prevalence coefficients of the topic. From equation 2.7, we can see that a regularizing prior centered around the zero vector on each column of $\boldsymbol{\Gamma}$. Note that STM only

models the prevalence of the first topic $K-1$, as the prevalence of the last topic, topic K, is known once the values of the first $K-1$ are estimated. The matrix $\mathbf{\Sigma}$ is a $K-1 \times K-1$ variance/covariance matrix representing the correlation between the prevalence of the topic. Together, these parameters and covariates influence the proportion of a document devoted to a given topic and which topics are likely to appear in the same documents. For example, if one of the covariates under consideration is the date of publication and we are considering a corpus of newspaper articles that contain a topic of the National Football League (NFL), then we would expect a higher prevalence of this topic on dates that correspond to Mondays as compared to dates that correspond to Thursdays. The difference in topic prevalence is due to the nature of the NFL, where games are usually played on Sundays, meaning that there is more to report on the NFL in the Monday paper vs. the Thursday paper.

In addition to topic prevalence covariates, STM allows for topic content covariates such as publication source and other metadata. The matrix Y represents the topic prevalence covariates and is of size $D \times A$; the rows of this matrix, y_d, are used in the generative model for STM. In STM, these covariates adjust the word distribution for each topic; for example, two publications might have slightly different word distributions for the same topic. The topic distributions over words are modeled using m_v, $\kappa^{(t)}_{k,v}$, $\kappa^{(c)}_{y_d,v}$, and $\kappa^{(i)}_{y_c,k,v}$. The value m_v represents the logarithmic transformation of the probability of the baseline word. For example, this value is similar to the overall mean in an ANOVA. The kappa values represent the logarithmic rate deviations from this baseline. The value $\kappa^{(t)}$ is a matrix $K \times V$ that contains the deviations at the topic level of the baseline, $\kappa^{(c)}$ is a matrix of size $A \times V$ that contains the devia-

tions for each level of the A different covariates. Finally, $\kappa^{(i)}$ is a three-dimensional array of size $A \times K \times V$ that accounts for any interaction between the topics and the covariates. If the topic content covariates are not included in the model, then only the values of m_v and $\kappa^{(t)}$ are included. As mentioned above, this setup is comparable to an ANOVA model with categorical variables, and this analogy continues because we cannot estimate both m_v and $\kappa^{(t)}$ separately. However, we can estimate $m_v + \kappa^{(t)}_{k,v}$ for each topic k. This allows us to estimate a $K \times V$ matrix of topic distributions, over the vocabulary provided. As an extension to STM, the authors describe placing sparsity-inducing priors, such as the Laplace prior [FHT10], on the κ values. The sparsity in these log-transformed rate deviations aids in the interpretation of the topics.

The STM plate diagram is shown in figure 2.7 and helps to illustrate the evolution of STM from LDA compared to the LDA plate diagram in figure 2.4. The core content model is still the same; for each word in a document, a latent variable identifies the topic distribution from which the word is drawn. The evolution comes from allowing the covariates X to influence the proportions of a given document to the topic and the covariates Y to influence the distribution of the topic in the vocabulary.

2.2 Sentiment Analysis

Sentiment analysis, also called opinion mining, is a field of study devoted to developing an automated process to analyze text to classify the author's emotional tone, attitude, or sentiment as positive, negative, or neutral. Sentiment analysis is

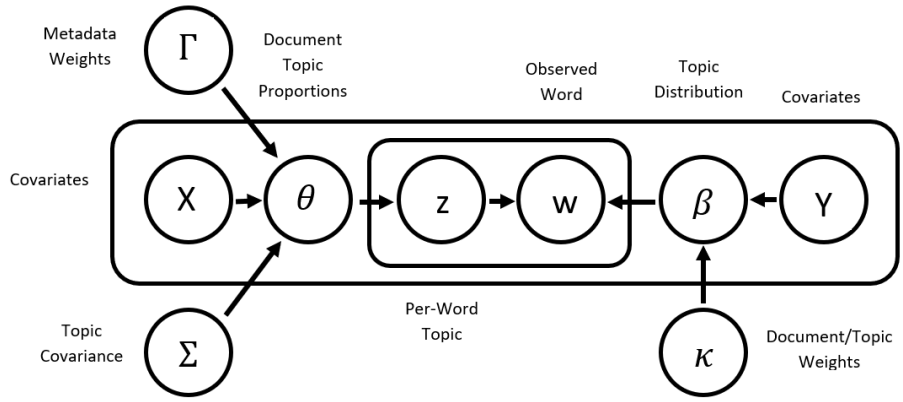


Figure 2.7: STM Plate Diagram. As with the LDA model, each plate represents a replicate. The values outside the plates represent the corpus level parameters, the values inside the first plate represent the document level parameters, and the values inside the inner plate represent the word level values within each document.

a rapidly growing field with contributions from both linguists and computer scientists. Sentiment analyzers have been utilized to analyze consumer attitudes towards purchased goods, entertainment such as movies, television shows and books, live events, and online reviews [LCT21]. Furthermore, sentiment analysis has become a valuable tool in understanding and interpreting social connections and interactions via social networks [Poz+17]. Currently, there are several surveys and overviews of the current state of the Sentiment Analysis field; see [WRK22], [PMS17], [You+19], and [BKB21].

The tools that analyze the text and classify the author's overall tone as positive, negative, or neutral, broadly speaking, come in one of two varieties. Either a machine learning algorithm, such as a neural network, Bayesian network, or support vector machine, is trained using a large labeled data set, or the alternative approach uses a dictionary of known words and their sentiment along with rule-based heuristics to

evaluate the sentiment of a document [WRK22]. A third option exists, which is a hybrid of the ML and lexicon approaches, but this is less common and requires the tools of both approaches. The use of machine learning tools requires a large training data set that contains a classification of each document in the training set as either positive, negative, or neutral. These tools work well in consumer product reviews and surveys, where a response is readily available and is often included with the text. Suppose that such a training set does not exist or would be prohibitively expensive to obtain. In that case, a lexicon sentiment analyzer is the only practical option, as these tools are based on a predefined dictionary and syntactical rules to determine the sentiment of a document or text.

This section briefly summarizes the current state of the machine learning models used in sentiment analysis and reviews the lexicon-based sentiment analyzer used in this project to define and quantify polarization.

2.2.1 Machine Learning Approaches

Machine learning (ML) is a phrase used to describe a broad collection of classification algorithms; that is, these algorithms focus on accurately predicting the category/-classification of an object (picture, text, etc.) based on a set of features. These features can be continuous or discrete. One hallmark of machine learning algorithms is the need for a large prelabeled training data set. The size of the training set varies depending on the algorithm, with some algorithms performing poorly if the training set is too large, such as Naïve Bayes (NB), however, most algorithms require a large

data set to reach peak performance.

The application of machine learning algorithms in sentiment analysis has increased dramatically in the last decade. For example, Kang et al. used a Naïve Bayes algorithm to classify restaurant reviews [KYH12]. Tripathy et al. implemented a Naïve Bayes algorithm and Support Vector Machine (SVM) algorithm to classify movie reviews as positive or negative [TAR15]. With the NB algorithm, they obtained an accuracy of 0.895, and with the SVM algorithm, they obtained an accuracy of 0.94. Chen and Tseng used two varieties of multilevel class SVM algorithms, One-Versus-All SVM and Single-Machine Multiclass SVM, to classify reviews from a variety of product areas, see [CT11]. Additionally, Jain et al. implemented various machine learning algorithms, including Decision Trees (DT), to distinguish between genuine and fraudulent reviews of hotels and restaurants [JPA21]. The works cited here are just the tip of the iceberg in terms of the uses and applications of machine learning algorithms to sentiment analysis. However, each of these uses has one thing in common: they all require domain-specific training data sets that are appropriately labeled. This requirement is why many applications for these algorithms center around areas where a prelabeled data set, such as product reviews, is easily obtainable.

Given that our area of interest is the domain of news articles, an easily available data set of articles labeled as positive, negative, or neutral does not exist at this time. Given this limitation, we focus on lexicon-based sentiment analyzers, which do not require a large training data set and are often called unsupervised learners.

2.2.2 Valence Aware Dictionary for Sentiment Reasoning

The Valence Aware Dictionary for sEntiment Reasoning (VADER) is a lexicon based sentiment analyser developed by Hutto and Gilbert [HG14]. As mentioned above, VADER is a rule-based lexicon-driven sentiment analyzer initially developed to analyze the overall tone of social media posts. However, the authors claim that it also "readily generalizes to multiple domains". In order to create VADER, the authors had two primary tasks: 1) "the development and validation of a gold standard sentiment lexicon that is sensitive to both the polarity and the intensity of sentiments expressed" and 2) "the identification and subsequent experimental evaluation of generalizable rules regarding conventional uses of grammatical and syntactical aspects of text for assessing sentiment intensity". The necessity of step 1 is obvious, and the necessity of step 2 becomes evident once one considers sentences such as "Overall the food was good, but the appetizers were not the best." In this sentence, two words with positive sentiment appear "good" and "best," however, only "good" is used with positive sentiment given the full context of the sentence. The "not" two words before "best" negate the sentiment of "best," flipping the overall sentiment to negative. Thus, having a lexicon of sentiment-laden words and their corresponding sentiment score is not enough to accurately qualify the sentiment of a complete sentence.

In order to develop their lexicon, Hutto and Gilbert opted for a human-centered approach to construct and validate their sentiment lexicon. First, they used well-established sentiment word banks to populate their lexicon list and added additional words, acronyms, and western-style emoticons. In total, they had 9,000 lexical fea-

tures for consideration. In order to assess these features, the authors utilized Amazon Mechanical Turk (AMT), a platform where workers can perform minor tasks for small amounts of money. These workers rated each lexical feature on a Likert scale, with -4 representing the most negative sentiment a feature could have, a 4 representing the most positive sentiment a feature could have, and a 0 representing neutral or no sentiment. After evaluating all characteristics, the authors kept any characteristic that did not have a zero mean sentiment value and had a standard deviation less than 2.5. This left a total of slightly more than $7,500$ lexical features. As examples, words such as "okay" have a value of 0.9, "good" is 1.9, "great" is 3.1, "horrible" is -2.5, and "sucking" is -1.5.

With their "gold-standard" lexicon in place, Hutto and Gilbert set out to develop heuristics that would allow them to apply their lexicon to fully formed sentences and extract the observable sentiment in the sentence. To determine these heuristics, the authors selected 800 sentiment-laden social media posts, 400 with extreme positive sentiment and 400 with extreme negative sentiment. The experts then rated each post and rated them on the same scale -4 to 4 described in the previous paragraph. The authors then compared this with the expected sentiment based on the developed lexicon. Hutto and Gilbert devised five heuristics that they employed as part of their sentiment analysis. The five heuristics quoted directly from the manuscript are the following.

1. "Punctuation, namely the exclamation point (!), increases the magnitude of the intensity without modifying the semantic orientation. For example, "*The*

food here is good!!!" is more intense that "*The food here is good.*"

2. Capitalization, specifically using ALL-Caps to emphasize a sentiment-relevant word in the presence of other non-capitalized words, incrases the magnitude of the sentiment intensity without affecting the semantic orientation. For example, "*The food here is GREAT!*" conveys more intensity that "*The food here is great!*"

3. Degree modifiers (also called intensifiers, boosters words, or degree adverbs) impact sentiment intensity by either increasing or decreasing the intensity. For example, "*The service here is extremely good*" is more intense than "*The service here is good*", whereas "*The service her is marginally good*" reduces the intensity.

4. The contrastive conjuction "*but*" signals a shift in sentiment polarity, with the sentiment of the text following the conjunction being dominant. "*The food here is great, but the service is horrible*" has mixed sentiment, with the latter half dictating the overall rating.

5. By examining the tri-gram preceding a sentiment-laden lexical feature, we catch nearly 90% of cases where negation flips the polarity of the text. A negated sentence would be "*The food here isn't really all that great.*" [HG14]

After defining these heuristics, the authors selected 30 social media posts and modified them using these heuristics. In total, they generated 200 posts and again used AMT workers to evaluate the sentiment of each post. The results show a statistically

significant difference in the sentiment rating for the posts altered by one of the above heuristics. The magnitude of the intensity shift is different for each heuristic, but the authors obtained a point estimate for each shift, which they utilized to implement their sentiment analyzer.

Although Hutto and Gilbert intended VADER to analyze the sentiment of social media posts, they also tested VADER in other domains, including NY Times editorials. They used 20 trained AMT workers to obtain the "ground truth" of the sentiment for this corpus. Then, they compared the results of VADER with other sentiment analyzers and even additional human interpretations. In the area of New York Times editorials, VADER outperformed every other automated sentiment analyzer and compared well to the additional human interpretations. The human raters obtained a precision of 0.87 and a recall of 0.55, which gives an F1 score of 0.65. VADER obtained a precision of 0.69 and a recall of 0.49, giving an F1 score of 0.55.

Thus, given that VADER is an unsupervised sentiment learner and the performance of VADER on a corpus similar to our domain of interest, we conclude that VADER is an acceptable tool to analyze the sentiment of our news data. In Chapter 3, we will detail how we implement VADER in our corpus and how we utilize the resulting sentiment analysis, along with topic modeling, to generate polarization plots, which we then use to quantify changes in polarization.

Chapter 3

Defining Polarization

Many fields use the word polarization to define a situation or scenario where, instead of having a continuum of possible outcomes, only a few, usually two, possibilities on opposite extremes are observed. For example, in physics, unpolarized light becomes polarized when passed through a filter that blocks all light except light with an electric component oscillating along the filter's polarization axis. Figure 3.1 shows a diagram of this physical phenomenon. For this project, we define and utilize polarization in a sociological context. In this way, we rely on the work of McCoy et al. to help us define polarization and provide insight into how polarization can be quantified. To that end, McCoy et al. define polarization as "a process by which the normal multiplicity of differences in a society increasingly aligns along a single dimension, cross-cutting differences become instead reinforcing, and people increasingly perceive and describe politics and society in terms of "Us" versus "Them" [MRS18]."

Applying this definition to the idea of quantifying polarization to news sources, one would anticipate a state of high polarization to be one where the "single dimension" mentioned in the definition defines how events discussed in the news are described

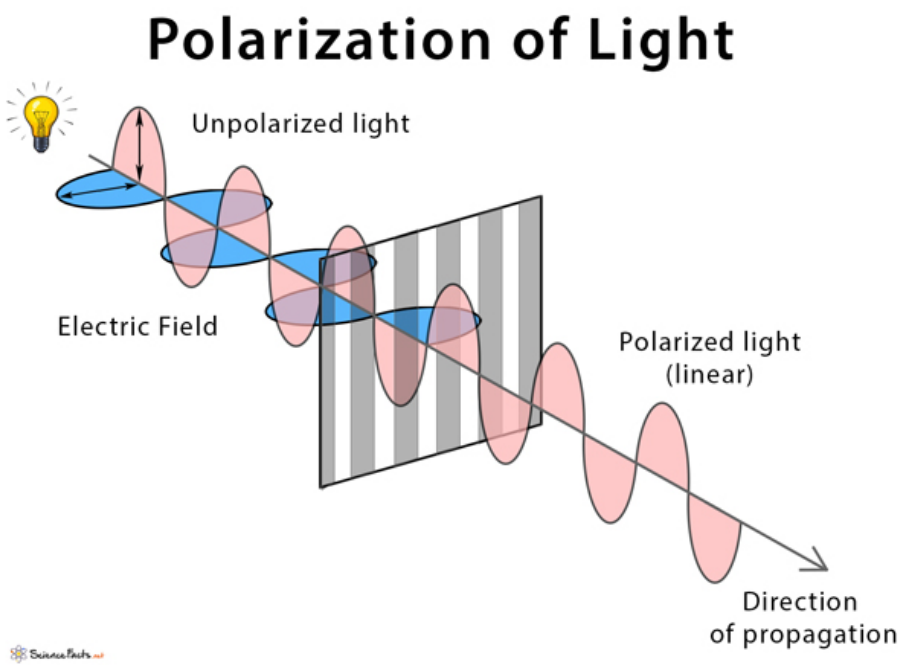


Figure 3.1: Polarization of light along an axis.
[Fac]

and presented to an audience. In this scenario, the majority of newsworthy events are presented, or "filtered" to extend the metaphor of polarized light in such a way as to align with one direction of that single dimension. Our goal in this work is not to describe or characterize the filter or the filter's dimensions. Instead, we look to define a process by which changes in the magnitude or alignment of filtering can be measured and compared to perceived or anecdotal changes in societal polarization. However, having the analogy of polarized light is useful when thinking about polarization in the news. When light is polarized by a filter, as demonstrated in figure 3.2, the only light rays entirely obscured by the filter are the rays with electric fields perpendicular to the filter's axis. In this way, all other rays have some component of their electric field projected onto the filter's axis. Applying this analogy to polarization in the news, we can think of a highly polarized environment as one in which various news topics are presented and described in part by their projection onto this polarizing axis created by the polarizing environment. That is, the most polarizing aspects of culture will influence and alter the discussion around all topics, even those that are marginally related to the polarizing aspects. To extend the analogy of polarized light even further, this work does not describe the filter that creates the polarized light, or in our case, the polarized environment, how the filter works, or what creates the filter. We aim to develop a method to detect when topics become more or less polarized.

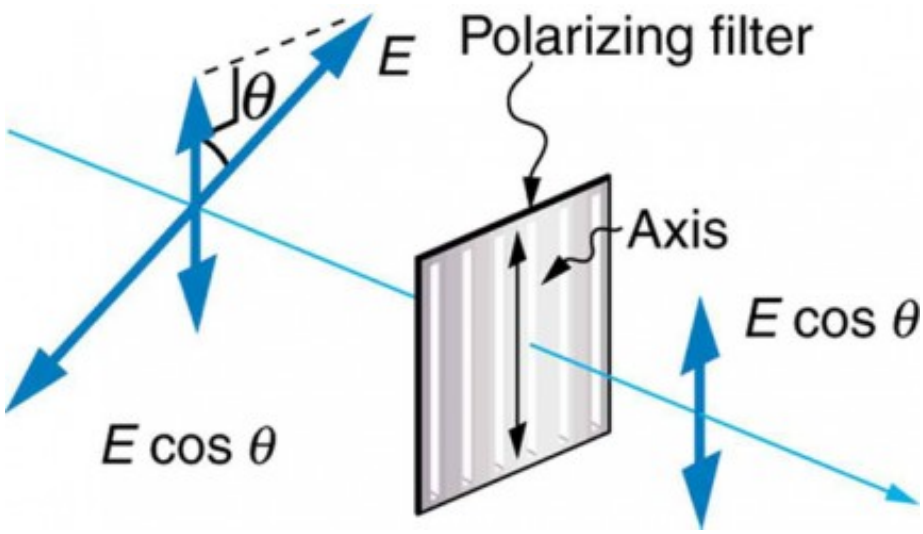


Figure 3.2: Polarization of light along an axis.
[Lea]

3.1 Polarization and Sentiment Distributions

Building off of this notion that a polarized environment filters the news in such a way as to align with one direction or the opposite direction along a social or political axis, we developed a method that combines both sentiment analysis and topic modeling that quantifies any disparity in the discussion of topics. Figure 3.3 gives an example of two topics that are not polarized and an example of a polarized topic. In the first two topic examples, we have instances where there is mainly agreement on the direction of sentiment used to discuss the topic. Topic 1 is chiefly discussed by authors using a generally positive sentiment, whereas those discussing topic two most often do so with a negative sentiment. In topic 3, we have an instance where there is general disagreement on the appropriate sentiment that authors and editors use when writing about topic 3. We argue that polarized topics will display this signature

of approximately a 50/50 distribution of positive to negative sentiment. Not all topics displaying this positive-to-negative sentiment distribution will be polarized or indicate overall polarization in society or news sources. Recall our polarized light analogy; when society is polarized, many topics will be filtered to align with the most polarizing issues of the day. A topic may have general disagreements about the discussion of the topic and what sentiment is appropriate without that topic being polarized or filtered based on a polarized environment. However, if a topic is polarized or presented in a polarized environment, we assert that it will have a polarization distribution similar to the one described in topic 3. One way to distinguish between a polarized topic and a non-polarized one with general societal or regional disagreement is to match topics over time and look for changes in this polarization distribution. In Chapter 5, we show specific examples of polarizing events and how changes in the polarization distribution correspond to known polarizing events or times with a heightened level of polarization.

3.2 Polarizing Events and Social Capital

This work on polarization is in part inspired by the concept of social capital, which "is a complex multidimensional concept that encompasses a repertoire of cultural and social value systems" [BY09]. While social capital is not a new concept, some claim the beginnings of what is now known as social capital go as far back as Karl Marx (1818-1883) and John Dewey (1859-1952), among others, scientific research into quantifying the impact of changes in social capital are at most a few decades

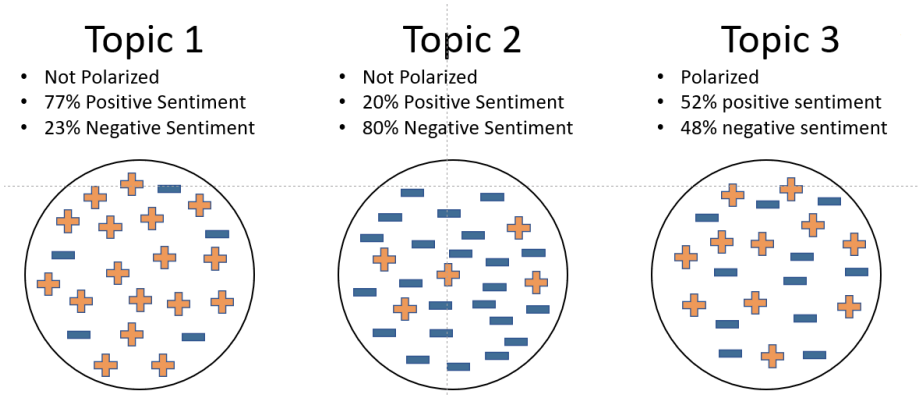


Figure 3.3: Examples of sentiment distribution for three topics: Topics 1 and 2 are not polarized, as overall agreement exists on the sentiment used to discuss the topic. Topic 3 is polarized, as roughly 50% of articles discuss the topic with a positive sentiment and half with a negative sentiment, indicating disagreement on how the topic should be discussed.

old. Bhandari and Yasunobu broadly define social capital as "a multidimensional phenomenon encompassing a stock of social norms, values, beliefs, trust, obligations, relationships, networks, friends, memberships, civic engagement, information flows, and institutions that foster cooperation and collective actions for mutual benefits and contribute to economic and social development." Clearly, social capital encompasses several factors, along with their complex interactions. In order to understand the role social capital has on polarization, we will focus on two aspects of social capital, bonding capital, and bridging capital.

Hawdon et al. [Haw+20] define bonding capital as "the trust of specific others", that is, people with whom we know and interact directly, such as family members, coworkers, and community members. Bond capital allows us to form and maintain social groups based on a shared belief, common goal, or cause. Conversely, bridging

capital extends social and cultural ties to persons outside one's direct social groups. Simply put, bonding capital creates and maintains groups that interact directly with each other, and bridging capital connects individuals across groups.

Given this context, we anticipate that changes in bonding and bridging capital will impact or cause changes in social polarization. For example, we hypothesize that social polarization will decrease when bridging capital is high, allowing people to connect and interact with other individuals beyond their direct social groups. On the other hand, we anticipate moments of low bridging capital and high bonding capital to correspond to moments of high social polarization, creating an "Us vs. Them" mentality. These insights guide us in what periods of time to look for high polarization and when to look for periods of low polarization. For example, we anticipate the weeks following the attacks of 9-11 to be a low-polarization moment. Obviously, the events of that day were tragic, but in the aftermath, the nation experienced a surge in bridging capital, allowing disparate groups to bond over the shared tragedy. Conversely, the ongoing war in Iraq during the mid-1990s-2000s could be a moment of high polarization, as the bridging capital between groups that supported the war and groups that opposed the war was low. We shall apply our understanding of bridging and bonding capital to help illuminate possible moments in time corresponding to high and low social polarization.

3.3 Ethics

In any statistical endeavor, ethical concerns about the work and potential impacts are of utmost importance, and one should take care to review ethical guidelines to ensure appropriate and ethical treatment of the data, stakeholders, fellow statisticians and anyone impacted by the decisions made based on statistical analysis. The American Statistical Association published ethical guidelines for statistical practice on February 1, 2022 [Pro] and outlines eight aspects that one should consider when practicing statistics. We present these eight facets below:

- Professional Integrity and Accountability
- Integrity of Data and Methods
- Responsibilities to Stakeholders
- Responsibilities to Research Subjects, Data Subjects, or Those Directly Affected by Statistical Practices
- Responsibilities to Members of Multidisciplinary Teams
- Responsibilities to Fellow Statistical Practitioners and the Profession
- Responsibilities of Leaders, Supervisors, and Mentors in Statistical Practice
- Responsibilities Regarding Potential Misconduct

We have followed these guidelines in all aspects of this work as faithfully and closely as possible. We have maintained our data integrity and have not altered any data pro-

vided by NewsBank; see Section 4.1. Although our overall method is novel, we have applied any existing statistical technique in a valid and appropriate manner, including discussing the limitations and shortcomings of the methods used. Specifically, for topic modeling, see Section 2.1. We have worked to ensure that all stakeholders, fellow statistical practitioners, and members of our multidisciplinary team have been prioritized and informed of any changes or results that could affect them. Finally, we have taken great care to ensure that any conclusions drawn from this work are sound and well supported by the data.

Chapter 4

Quantifying Polarization

With our definition of polarization in place, background on topic modeling, and sentiment analysis, we are now ready to outline our process for quantifying polarization. There are four main steps: acquire a representative sample of news articles that contain polarizing events as well as nonpolarizing environments, apply the sentiment analyzer to each article to quantify the overall sentiment of the article, use a topic model to extract the topics out of the corpus, and finally combine the sentiment scores of documents that contain the same topic to generate polarization plots. Figure 4.1 outlines the process described above. In the following sections, we detail each step in this process.

4.1 Data Source

In order to test our method, we selected NewsBank, a curated database of newspaper articles dating back more than 100 years. NewsBank offers a variety of publications, from local to national newspapers. We set out to get a representative sample in many dimensions. The dimensions we considered were political views/leanings, if known,

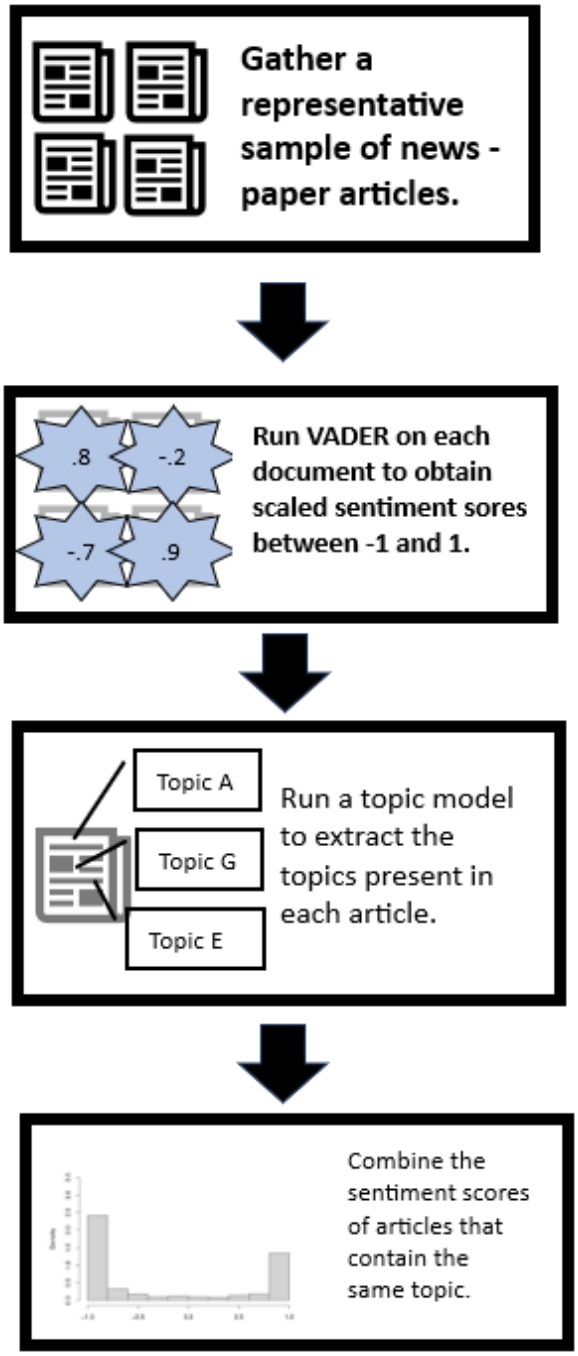


Figure 4.1: Quantifying Polarization Process Outline.

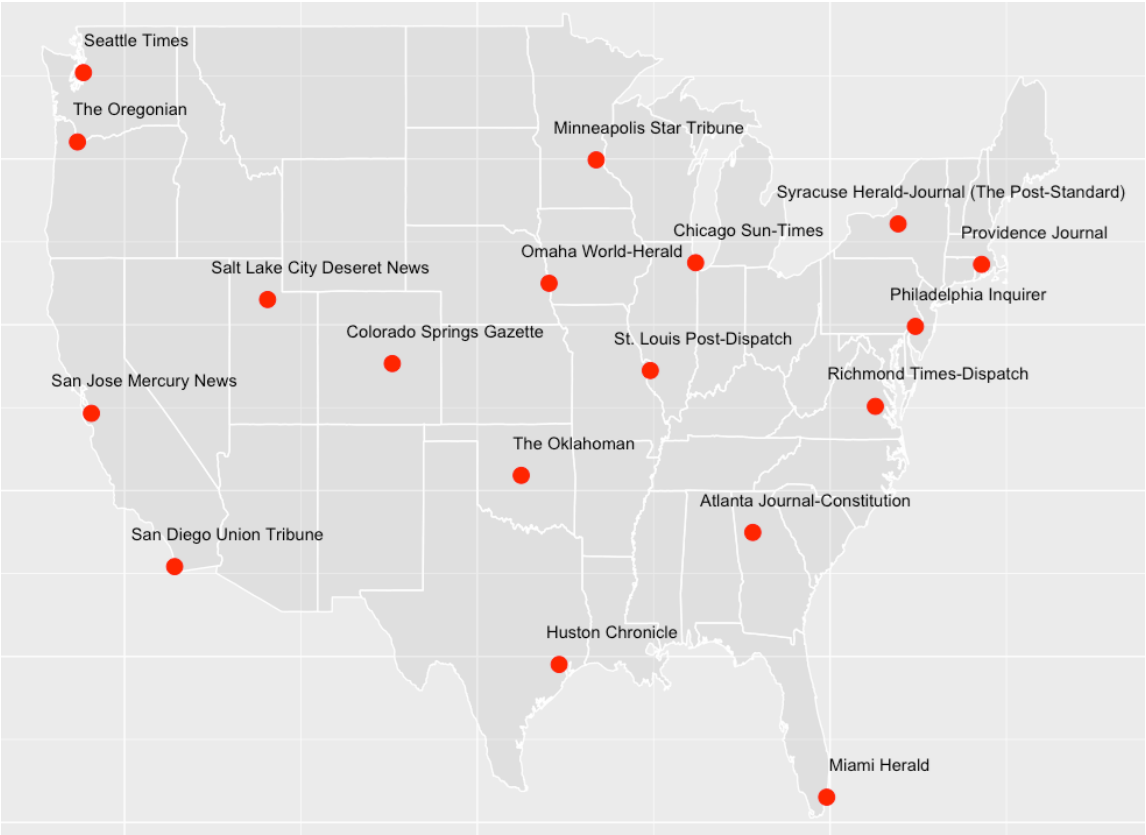


Figure 4.2: Location and name of the eighteen regional papers included in the NewsBank data set.

location of headquarters, and national vs. regional distribution base. We also had a minimum requirement that the publication contain both a variety of sections and news events. This requirement eliminates publications that focus only on one news area, such as politics or sports. In total 20 publications were selected, 18 regional papers and 2 national papers. The two national papers are USA Today and Christian Science Monitor, and the 18 regional papers are shown in Figure 4.2 which displays a map of the United States, as well as the name and location of each regional paper.

In addition to the articles themselves, NewsBank maintains additional metadata on the articles. This metadata includes author, display date, headline, section, publisher, and several internal NewsBank variables. Given that the topic model, Structural Topic Model (STM), used in the analysis allows for the inclusion of covariates, our original hope was that this metadata would be helpful to include in the analysis. However, much of the metadata is too sparse to be analyzed; for example, there are not enough articles by each author for a meaningful result. Additionally, labels are inconsistent between publications; for example, not all sections have consistent labeling across each of the publications to make comparisons meaningful. The publication date and name are the only metadata that consistently contains values and the repetition needed for analysis.

In addition to obtaining a representative sample of articles based on the publication source (location, national vs. local, etc.), we wanted to include periods of anticipated polarized environments and periods of anticipated low or non-polarized environments. Examples of anticipated polarized environments include regularly occurring events such as presidential elections as well as one-time events such as the beating of Rodney King by Los Angeles police and the subsequent riots that lasted for six days following the acquittal of the four police officers involved in the incident. Examples of anticipated environments that are low or non-polarized are national sporting events, such as the summer and winter Olympics, and the immediate aftermath of the 911 terrorist attacks. Examples of each environment are given in Table 4.1.

In order to capture the lead-up and subsequent coverage of each event, news starting

Highly Polarized Environments	Low Polarized Environments
Pulse Nightclub Shooting June 12, 2016	Summer Olympics 1988, 1992, 1996, 2000, 2004, 2008, 2012, 2016, 2021
Presidential Elections 1988, 1992, 1996, 2000, 2004, 2008, 2012, 2016, 2020	Winter Olympics 1988, 1992, 1994, 1998, 2002, 2006, 2010, 2014, 2018
Death of George Floyd May 25, 2020	VirGinia Tech Shooting April 16, 2007
Cambridge Analytica Controversy 2013-2017	Oklahoma City Bombing April 19, 1995
L.A. Riots Apr 29, - May 4, 1992	Direct Aftermath of 911 Sept 11 - 30, 2001

Table 4.1: Table of Anticipated Polarizing Events and Environments and Anticipated moments of low Polarization.

one to two months before the event, as well as one to two months after the conclusion of the event, were included in the corpus of selected articles. Having this allows us to baseline how polarized the environment was before the event and how the environment changed after the event. In total, 119 months of news articles between 1988 and 2021 are in the data set.

4.2 Sentiment Scores

Along with the publication of the VADER sentiment analyzer [HG14], the authors created a Python package, vaderSentiment 3.3.2, based on their work to analyze text. This package allows the user to create a sentiment intensity analyzer object that takes in a string of text and returns four values of interest, a percentage of the text that is positive, negative and neutral, and a scaled combination of these numbers that the authors of VADER call the compound score. These compound scores are

the sum of the sentiment scores based on the heuristics discovered by the authors of VADER and the developed sentiment lexicon. The formula for the compound score is shown in Equation 4.1 below, where *sent_sum* represents the sum of the sentiment scores described above, and α is a normalizing constant set to 15 by default in the VADER package.

$$compound\_score = \frac{sent\_sum}{\sqrt{sent\_sum^2 + \alpha}} \tag{4.1}$$

From the above equation, it is clear that the compound score is scaled to be between -1 and 1, where -1 indicates that the text has an exceptionally negative sentiment and 1 indicates that the text has an extremely positive sentiment. These compound scores are the values computed in step two of Figure 4.1 and create the polarization distributions in step four of the exact figure. By default, the α value in the compound score computation is not changeable in the original Python code provided in the vaderSentiment 3.3.2 package. However, since the above function to compute the compound score is invertible given α, it is possible to compute the value *sent_sum* for each article and recompute $compound\_score_\alpha$. Note that for notational purposes, we have added an α to the subscript to reflect that *coumpound_score* now depends on the value of alpha chosen.

4.3 Topic Modeling

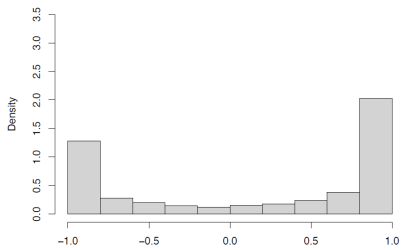
The topic model used in our analysis is the Structural Topic Model (STM). The model was fitted using the R package developed by the authors [RST19]. We included several covariates in the analysis based on the metadata provided by NewsBank and the VADER output. For most of the analysis, the covariates included in the NewsBank metadata are the publication date, the name, and the newspaper section. From the VADER output, the compound score was also included as a covariate.

The sheer volume of articles considered in this analysis prohibits fitting any topic model to the entire corpus. Thus, a reasonable method of subsetting the data is required to appropriately fit the STM in an interpretable manner and to evaluate the polarization of the resulting topics. Given that the initial thrust of this project centered on polarizing events/environments compared to moments of low polarization, it was a natural fit to analyze the data surrounding those events/periods. As we shall see in Chapter 5, the overall difference in sentiment, and thus polarization, between these events was minimal. Thus, a new approach was needed. We decided to gather all the data that we could access and group the articles on a monthly scale. This grouping of the articles allows us to obtain solid estimates of the main topics of discussion over a more extended period while also allowing for enough observations to monitor how topics change over time.

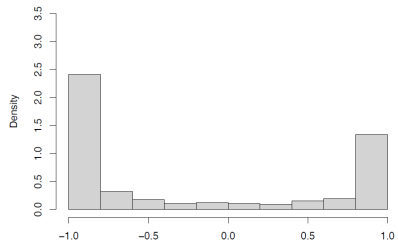
4.4 Polarization Distributions

Given our definition of polarization from Section 3.1, we now combine the output of sentiment analysis and topic modeling to generate polarization plots that can be studied and analyzed to quantify polarization. According to our intuitive definition of polarization, illustrated in figure 3.3, we need to match documents based on the topics contained within the document. Since STM does not produce sparsity in the topic/document proportion estimates, each document will likely contain at least a small percentage of each topic according to the estimated fit. Given this, an inclusion criterion for a topic/document proportion will determine which documents "belong" to a given topic and which do not. Arguments exist for a variety of values for such a threshold; however, without a method of estimating such a quantity, we are left with the typical hyperparameter estimation method, such as cross-validation. In chapter 5, we will explore the effects of this hyperparameter on the conclusions reached and the effects that such a value can have on the polarization distributions. We shall proceed knowing that such a quantity is needed. However, we shall momentarily forego assigning a specific value to such a quantity.

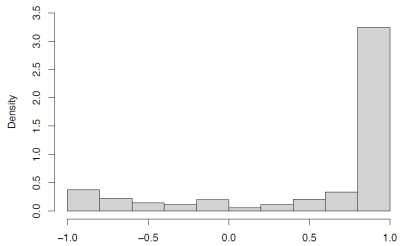
Once the documents have been aligned based on the topics they contain, we can create a distribution of compound scores for a given topic, which we call *polarization distributions*. Figure 4.3 gives examples of such polarization distributions. An interesting note about each of these distributions is that the majority of the density is around the extreme values of -1 or 1, with all four distributions having a very low density for the value between -1 and 1. This distribution and shape of the



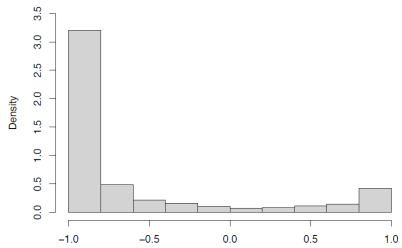
(a) Example Topic with Slightly Positive Sentiment



(b) Example Topic with Slightly Negative Sentiment



(c) Example Topic with Majority Positive Sentiment



(d) Example Topic with Majority Negative Sentiment

Figure 4.3: Examples of Polarization Distributions for polarized topics, 4.3a and 4.3b, and non-polarized topics, 4.3c and 4.3d
.

polarization plots is a common occurrence, as we shall see in Chapter 5.

In figure 4.3, we have examples of four topics, each with varying degrees of polarization. Figures 4.3a and 4.3b show examples of polarized topics. The overall sentiment is slightly positive in figure 4.3a. However, there is still some disagreement on how to discuss the topic, with many documents using negative sentiments. A similar statement is true for figure 4.3b but with swapped negative and positive sentiment scores. In this topic, the general sentiment is slightly negative, but there are still many articles with positive sentiment. We defined this disagreement between authors on which sentiment is appropriate to use when discussing the topic as an indication of

a polarized topic in Section 3.1. Thus, we aim to use these polarization distributions as bases for quantifying polarization. We will use changes in these distributions to indicate either an increase or decrease in polarization. For example, in Figures 4.3c and 4.3d, we see examples of topics that are not polarized because there is general agreement on the sentiment used to discuss the topic. In the example of Figure 4.3c, the sentiment is overwhelmingly positive, and in Figure 4.3d, the sentiment is overwhelmingly negative. Thus, we have a general agreement on the tone used to discuss these topics. If a topic evolved from a polarization distribution of one figure such as 4.3c to that of a line figure 4.3a, we would argue that this topic has become more polarized. Alternatively, if the order of appearance of these distributions were reversed, we would argue that the topic has become less polarized.

As mentioned earlier, our objective was to quantify changes in these polarization distributions to detect periods of high or low polarization. Our original hope was that there would be apparent diversity among these distributions that would allow for the utility of various metrics that illustrate changes in polarization. Some examples of metrics that we considered are range, variance, skewness, percentage with positive sentiment, entropy, and Gini. As we shall see in Chapter 5, every polarization distribution that we investigated exhibited the same characteristics "U" shape shown in Figure 4.3. The only noticeable change in these distributions across topics appears to be the magnitude of the modes at 1 and -1.

Given that the distribution is bounded and that most articles have a compound score of 1 or -1, the range of these distributions changes very little between topics. Although variance and skewness exhibit more significant changes in distributions,

as we shall see in Chapter 5, the changes in these metrics were not substantial and did not correspond to the anticipated periods of changes in polarization. As shown in Section 4.2, it would be possible to modify the compound score for a given article, thus changing the polarization distributions and, in turn, possibly changing the variance or skewness values obtained from the polarization distributions by adjusting the α value used in the computation of the compound score. The justification for introducing such a value to standardize the sentiment scores is that α is the expected number of sentiment-laden words in a document. With no such value of α readily available for our purposes, we are left with two options: either attempt to estimate and appropriate α value from the data we have at hand or use metrics independent of α. Estimating the value α would require extreme caution to avoid the appearance of confirmation bias, as it could appear that the justification for changing the original value would be to obtain polarization distributions that were closer to our expectations. These metrics, variance and skewness, have the additional drawback that they are not intuitive and understandable to a general audience.

Percent positive sentiment offers an intuitive and interpretable metric for quantifying polarization and is completely independent of the choice of α. Of course, the metrics, entropy and Gini, share the property that both are independent of a choice in α, as both are functions of percent positive sentiment. However, they are less intuitive and interpretable than percent positive sentiment. In Chapter 5, we shall compare these potential metrics and their performance related to quantifying polarization.

4.5 Post Processing

Computing the polarization distributions for each topic in our data set is insufficient to make any meaningful discoveries or interpretations. At this point, all we have are $2,380$ polarization distributions without understanding which topics these distributions represent and how these distributions might change over time. In order to remedy this issue, we need to place a human interpretable label on each topic and develop a method to match topic. These are the post-processing steps described in this section that allow interpretation and investigation of anticipated polarizing events.

4.5.1 Topic Labeling

As discussed in Section 2.1, topic models represent topics as probability distributions over words. This representation allows various methods to extract words that 11represent" a given topic. The most intuitive method is to examine the words with the highest probability of occurrence for a topic. Additional metrics exist that consider the exclusivity of a word to a given topic. For example, the FREX metric is the harmonic mean of the empirical CDF of a word for a given topic, along with the empirical CDF of exclusivity to that topic. The exact formulation of FREX is given in Equation 4.2.

$$\text{FREX}_{k,v} = \left(\frac{\omega}{\text{ECDF}(\beta_{k,v} / \sum_{j=1}^{K} \beta_{j,v})} + \frac{1-\omega}{ECDF(\beta_{k,x})} \right)^{-1} \tag{4.2}$$

Highest Prob.	food, cook, cup, serv, egg, minut, restaur, can, fat, use
FREX	recip, chees, onion, tablespoon, sauc, teaspoon, chicken, calori, veget, milk
Lift	aroma, artichok, calori, fat-fre, garlic, vinegar, -quart, apricot, blender, boneless
Score	tablespoon, teaspoon, sauc, cup, onion, calori, recip, egg, garlic, fat

Table 4.2: Topic three from March 1999.

In this notation, $\beta_{k,v}$ represents the frequency of the word v in topic k, and ω is a user-defined weight that alters the balance between frequency and exclusivity. For this work, we only considered the FREX values when $\omega = 0.5$, which is the default for the stm package.

The stm package uses other metrics in addition to FREX. These include the Lift and Score metrics, which, like FREX, attempt to balance a term's frequency and exclusivity. Lift is the ratio of a term's frequency divided by the frequency of that same term in the other topics. Score uses a similar ratio, except that instead of frequency, it uses the log of frequency.

The stm package provides a function that presents the top terms according to the above-mentioned metrics: high probability, FREX, Lift, and Score. To demonstrate how topics are labeled, we provide these metrics for three topics from March 1999. We present these terms precisely as they appear in the stm package, with only topic numbers representing them.

All three topics have a unique internal theme. The exact label used for each topic is irrelevant as long as the label gives a reasonable interpretation of the topic and is

Highest Prob.	game, point, play, score, team, season, shot, second, first, rebound
FREX	kazan, laker, knick, rodman, hawk, rocket, iverson, nba, benigni, clipper
Lift	divac, gugliotta, hornacek, latrel, shaquill, stoudamir, abdurrahim, buechler, carlesimo, crotti
Score	kazan, nba, rebound, laker, benigni, coach, game, knick, iverson, rodman

Table 4.3: Topic five from March 1999.

Highest Prob.	said, polic, charg, offic, court, case, investig, two, car, counti
FREX	prosecutor, arrest, convict, prison, sentenc, juri, jail, polic, crimin, crime
Lift	feloni, mistrial, sodomi, burglari, inmat, kevorkian, parol, robberi, youk, amadou
Score	polic, prosecutor, arrest, kevorkian, investig, attorney, murder, said, prison, crime

Table 4.4: Topic thirteen from March 1999.

consistent with other labels. For example, the topic presented in Table 4.2 is related to food preparation and possibly restaurant reviews. Thus, we labeled this topic "Cooking/Food." Even the most casual of sports fans will be able to recognize many of the terms in topic five as relating to the National Basketball Association, the NBA for short. The list includes several team names as well as the names of prominent players at the time. This topic was labeled "NBA." Finally, topic thirteen has several terms directly related to law enforcement and the legal system. We labeled this topic "Crime."

In order for our topic matching to be useful, we need to label all $2,380$ topics restricted over the 119 months worth of NewsBank articles. These labels will be referenced throughout the remainder of this document to distinguish topics. Once we can match topics, we will see that matched topics generally have consistent topic labels.

4.5.2 Matching Topics

Given the sheer volume of news material generated daily, weekly, monthly, and yearly, it would be computationally impossible to fit a topic model to any corpus of news articles that included diverse sources and spanned a considerable time. This constraint and the desire to observe changes in the polarization distribution of topics over time meant that we needed to develop a method for matching topics over time.

Because topics are probability distributions over words, we need a way to compare the similarity of probability distributions. One problem with traditional methods of comparing probability distributions, such as Kullback-Leibler divergence, is that

we do not have a guarantee that these probability distributions will have a common support. These topics span different corpora, making it highly likely that they will have different words as support. In order to obtain a common support, we select the top one hundred words with the highest probability for each topic and compare the associated probabilities. Of course, we must ensure that the words used to make the comparison appear in each topic. With this setup in place, we could compare topics based on only one hundred words if they have perfect agreement on the top one hundred most probable words and up to as many as two hundred if the topics have no words in common among their top one hundred.

The exact probability value associated with a specific word is less important than the relative order among the words for a given topic. Thus, once we obtain a common set of words to compare the topics, we compute the Kendall rank correlation coefficient, also known as Kendall's tau, between the two sets of probabilities. Kendall's tau measures the ordinal association between two associated vectors. We define two topics that are matched if their Kendall rank correlation coefficient is greater than zero and are not matched otherwise.

To illustrate the matching process, we present a scatter plot for matched topics and a scatter plot for unmatched topics. These plots are shown in Figures 4.4 and 4.5. Note that these plots use the logarithm of term probabilities.

In Figure 4.5, we see an "L" shape characteristic of unmatched topics. This "L" shape results from topics with very few words in common. Figure 4.4 shows a close linear relationship between the points. Again, this is an artifact of the two topics having

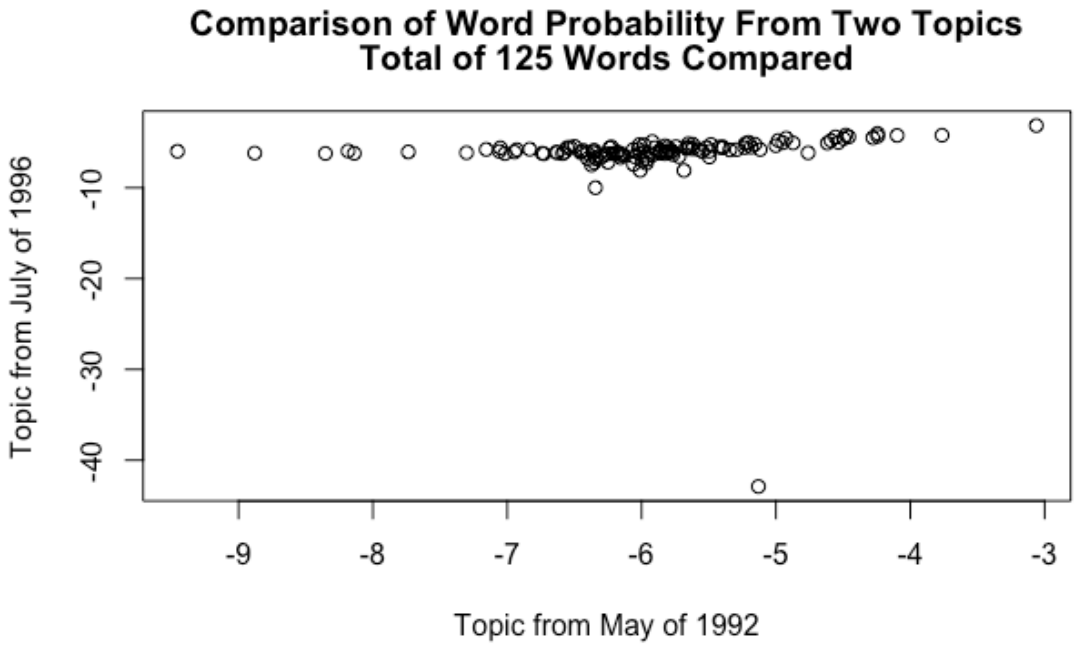


Figure 4.4: Scatter plot for matched topics.

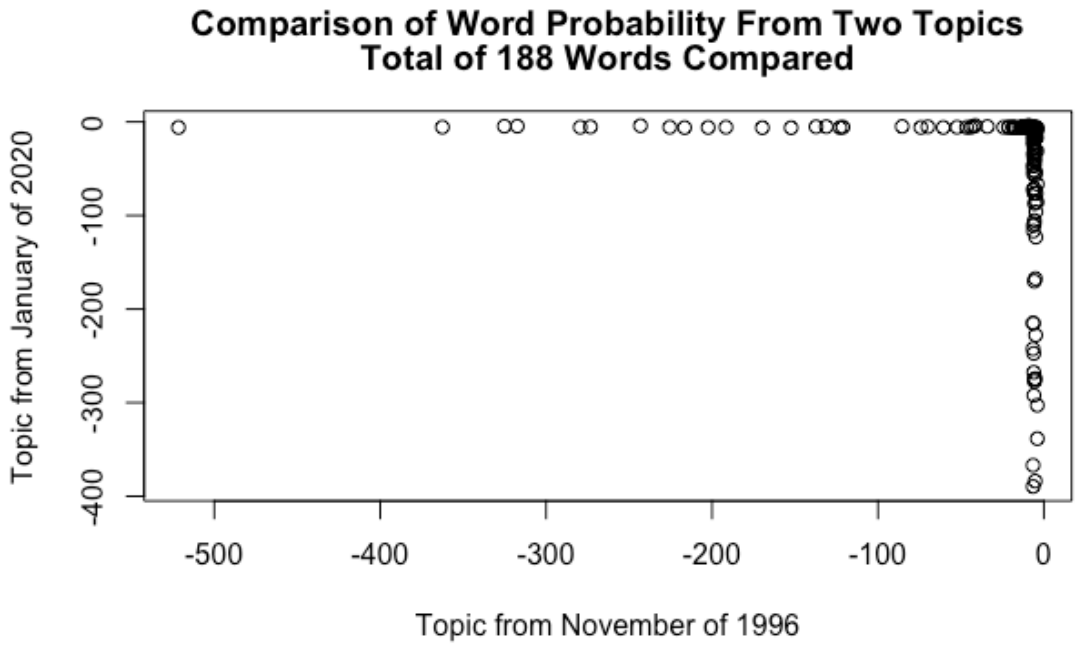


Figure 4.5: Scatter plot for unmatched topics.

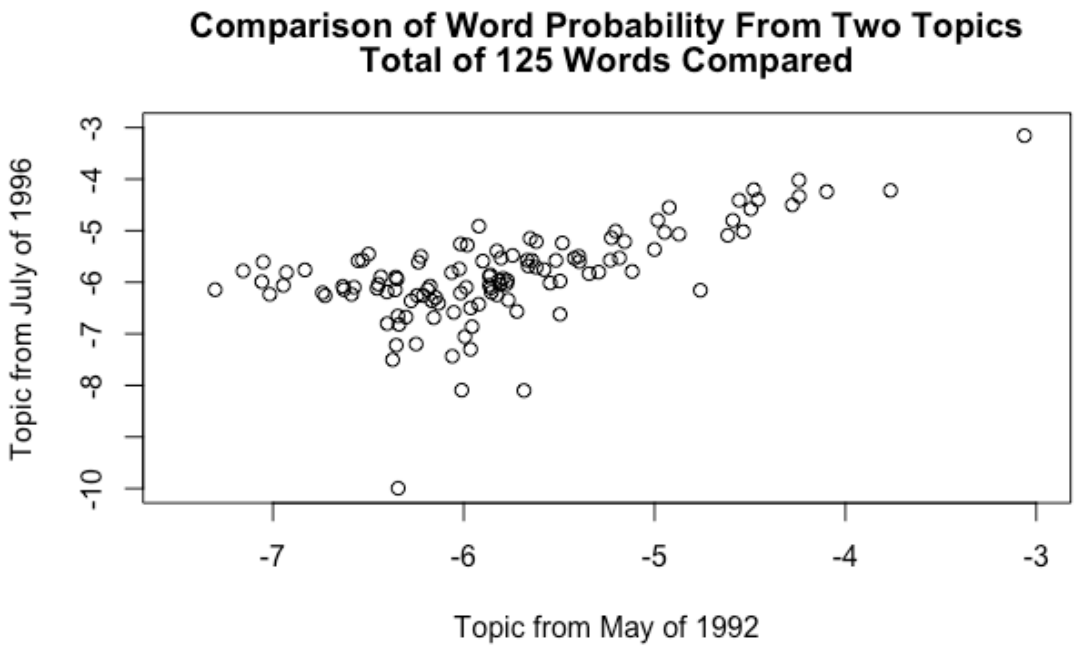


Figure 4.6: Zoomed in scatter plot for matched topics.

several words in common in their top 100 most probable words. To help illustrate this linear relationship, we zoom in on the cluster of points in the upper right section of the plot. Figure 4.6 shows this zoomed in scatter plot. It is clear from the plot that the topics have several words in common and a positive correlation between the term probabilities. These results suggest the question, why not use the total number of words in common instead of computing the Kendall rank correlation coefficient? While it is true that the number of words in common is related to Kendall's tau value, using the number of words in common has the downside of not providing a natural boundary to distinguish matched topics from unmatched ones. Using Kendall's tau provides this natural boundary of zero.

Matching topics over time allows us to compare polarization distributions and look for signatures of polarization for various topics. It also allows us to gain a historical perspective on the baseline polarization inherent to topics. We will present results in Chapter 5 based on a time series of matched topics.

Chapter 5

Results

This chapter is devoted to analyzing the polarization distributions obtained using the process described in the chapter 4. The presentation of this chapter is roughly the chronological order in which we investigated these events and their subsequent polarization plots.

5.1 Event Based Polarization

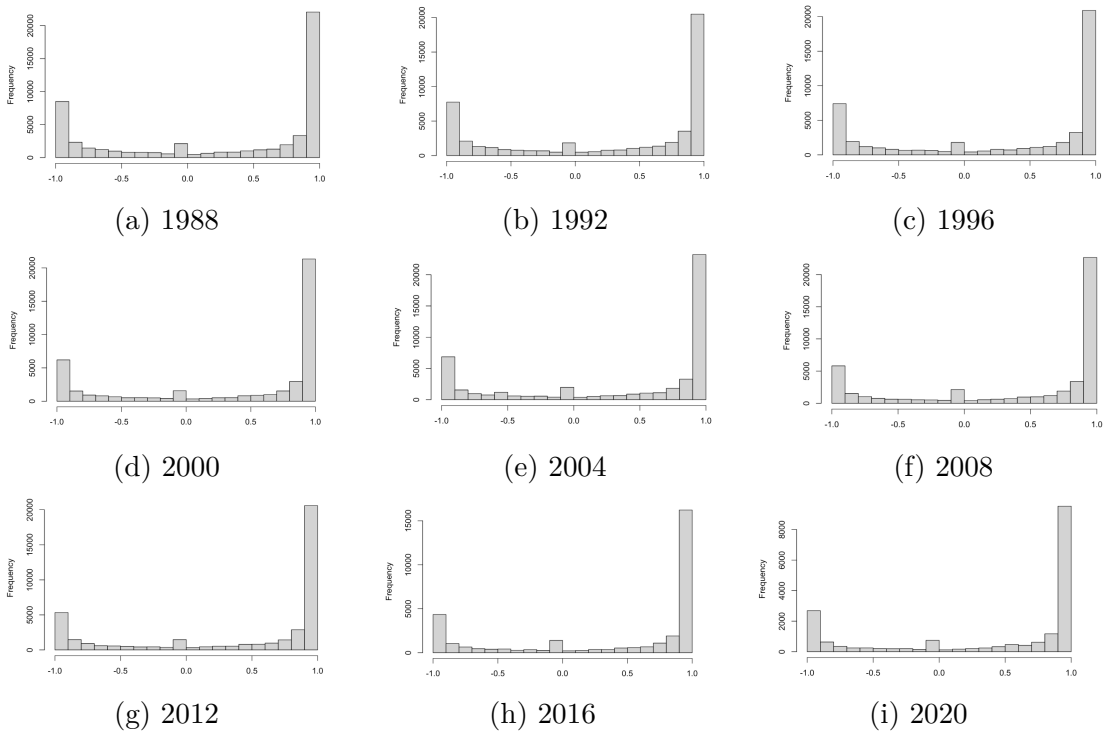
As mentioned in section 4.1, our initial hypothesis was that the degree of polarization would change based on specific events that have the power and impact to change the public discourse. For instance, we hypothesized that events such as the Los Angeles riots in the spring of 1992 would not only have a polarizing effect on the discourse, specifically relating to the riots, but could also have a polarizing effect on the discourse of events during the same period. We theorized that such events exacerbate the already present differences between social, political and economic groups, causing not only their differences in the riots to be illustrated in the discourse, but also the differences in discourse in other areas to become more pronounced. On

the other hand, we hypothesized that events such as 9-11, even given how terrible and tragic the events of that day were, would have an initial bonding effect, bringing the same social, political, and economic groups that might otherwise have been separated, together in unity, and thus a more extensive agreement on discourse. Our initial investigation of the NewsBank data explored the effects of such events.

5.1.1 Polarization Surrounding Elections

Political polarization is one of the most cited and studied sources of social polarization in academia [HL20]. Given the interest, one of the first events we investigated was to quantify the polarization around Presidential Elections in the United States of America. The NewsBank data contained articles surrounding each presidential election from 1988 to 2020. The US presidential election is statutorily set to be the first Tuesday following the first Monday of November, that is, the Tuesday between November 2 and November 8. In order to have a consistent basis from which to compare elections, we analyze all articles published ten days before the election and ten days after the election, including election day. We followed the process described in Chapter 4. To baseline the polarization plots obtained from various topics, we present the distributions of compound scores for all articles included in each time period. They are presented below in Figure 5.1

The first noticeable observation of these polarization plots in Figure 5.1 is how similar they are shaped over the years. All of these plots exhibit a somewhat "U" shaped distribution, but there are also distinctively three modes within each plot: the largest



(a) 1988 (b) 1992 (c) 1996

(d) 2000 (e) 2004 (f) 2008

(g) 2012 (h) 2016 (i) 2020

Figure 5.1: Polarization plots for all articles covering the 15 days preceding a US election, election day, and the 14 days following the election.

near 1, another more minor mode near -1, and finally, a slight mode near 0. As discussed in section 4.2, these values are the sum of the sentiment scores and a user-selected α value that the authors of the Python package characterize as "the maximum expected sentiment score for a given document." As mentioned earlier, VADER was designed to analyze social media posts, specifically Twitter (now X). We can assume that the authors expected 15 sentiment-laden words per tweet at most. With a set value of $\alpha = 15$, a document with an overall sentiment score greater than eight would have a corresponding compound score greater than 0.9. As we shall see later, this pattern of overwhelming extreme values in the polarization plots will be a recurring theme.

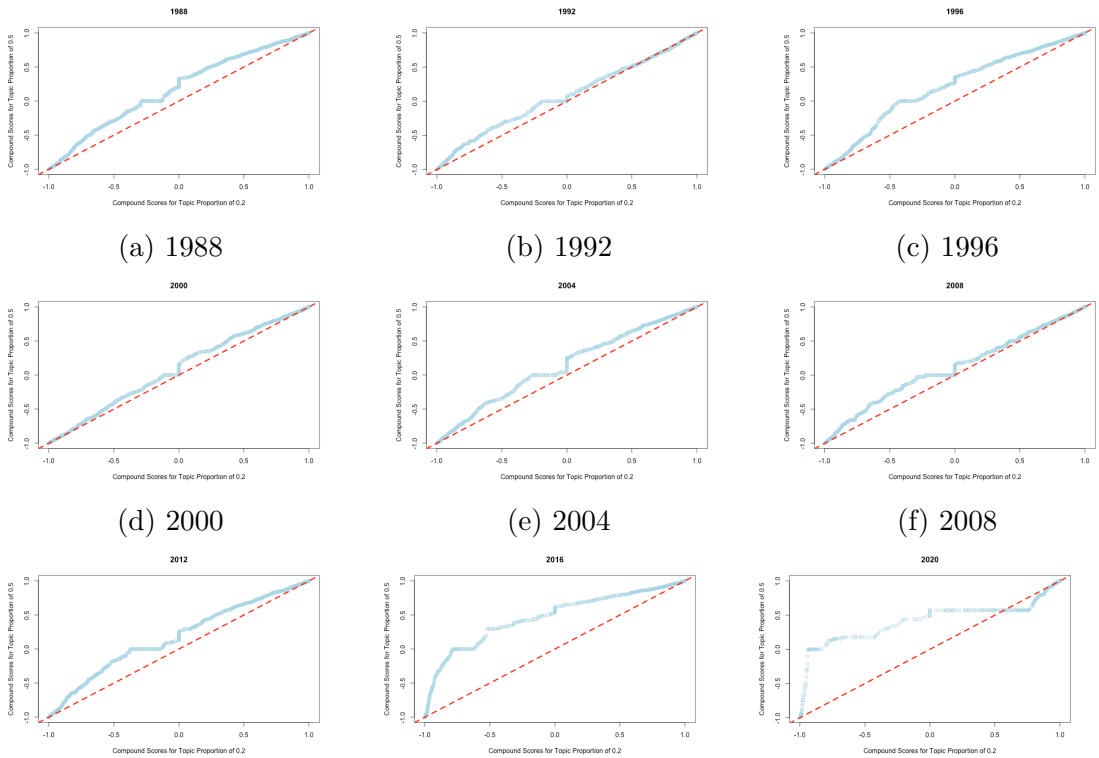
The second observation is that the scale for each of these graphs is the same except for the graphs for 2016 and 2020, as illustrated in Table 5.1. Each of these graphs covers the same number of days and even the same days of the weeks; however, there is a noticeable decline in the number of articles from 2012 to 2016 and from 2016 to 2020. We hypothesize that this reflects the general trend away from the print media [HL18]. The decline in the number of articles does not change the distribution of the overall polarization plots; it only decreases the magnitude of the frequency of each bin.

5.1.2 Polarization Plots for Topics Covered During the Presidential Elections 1988-2020

The STM topic model was fitted to the corresponding articles for each election year using $K = 20$ topics per year. Each topic was labeled using the technique described in Section 4.5.1, and the topic labeled "Presidential Election" was selected. However, we still needed to determine an appropriate value for the document/topic threshold used to classify a document as belonging to a given topic. As a point of comparison, we evaluated the compound scores, the values used in generating polarization plots, for the topics of the presidential elections from 1988 to 2020 using a threshold of 0.2 and 0.5. The resulting Quantile-Quantile (QQ) plots are shown in Figure 5.2.

The distributions for each of the years 1998 through 2012 are similar. The distribution created with a threshold of 0.5 skews slightly more positively than the distribution created with a threshold of 0.2. However, there are two years of Presidential Elections with very noticeable differences in the distribution of sentiment scores based on a threshold 0.2 versus a threshold 0.5, and these are the years of 2016 and 2020. These years have a very positive distribution with a threshold of 0.5. One potential explanation for such a difference is the number of articles included in each topic based on the threshold. These numbers are in Table 5.1. There is a noticeable drop in the number of articles in 1992, 2016, and 2020 compared to the other years.

One possible explanation for the drop in articles is that the "Election" topic was split during these years. We identify the topic of the Presidential election each year using



(a) 1988 (b) 1992 (c) 1996

(d) 2000 (e) 2004 (f) 2008

(g) 2012 (h) 2016 (i) 2020

Figure 5.2: Quantile-Quantile (QQ) plots of the compound scores based on a threshold of 0.2 and 0.5 for the Presidential Election topics for the years 1988 through 2020.

Threshold	1988	1992	1996	2000	2004	2008	2012	2016	2020
0.2	5862	3820	4738	4760	4563	4207	4956	3346	1994
0.5	2402	1179	1919	2135	1626	1693	2389	1141	784

Table 5.1: Number of articles included in the Presidential Elections topic based on Document/Topic proportion threshold.

the same techniques illustrated in Section 4.5.1. Regardless of the metric used to determine keywords for the topic, the names of the candidates are prominent. Each of the years in question has additional topics that are election-related. For example, in 1992, a topic labeled "Voting/General Election" was present with keywords such as "incumbent," "precinct," "voter," "republican," "democrat," and "district." Example headlines for this topic include "PEROT'S CAMPAIGN STYLE HAS INFLUENCE ON LOCAL RACE FOR STATE SENATE," "4 days until Nov. 3 election", and "BIG PUSH MADE TO GET VOTERS OUT". At a 0.2 threshold, this topic contains $3,820$ articles and $1,179$ articles at the 0.5 threshold. Given the similarities between these topics, it is likely that the Presidential Elections topic would contain many of these articles in other years, which could explain the discrepancy displayed in Table 5.1.

A similar issue appears in 2016 and 2020, with both years containing topics related to "Elections" in general and the anticipated aftermath of the Presidential Election. For example, 2016 contains a topic labeled "Election Aftermath/Social Unrest", which contains articles with headlines "Print only: Trump should commit to protective press pool for transparency," "Trump protests," and "Will Donald Trump require Muslims to register?" This topic contains $3,540$ articles at the 0.2 threshold and 618 at the 0.5 threshold. Again, it is likely that many of these articles would appear in the "Presidential Elections" topic in other years. The QQ plot for this topic at the two thresholds is shown in Figure 5.3. The compound score distribution is negatively skewed at the 0.5 threshold compared to the 0.2 threshold.

Given that, for most topics, there is very little difference between the distribution of

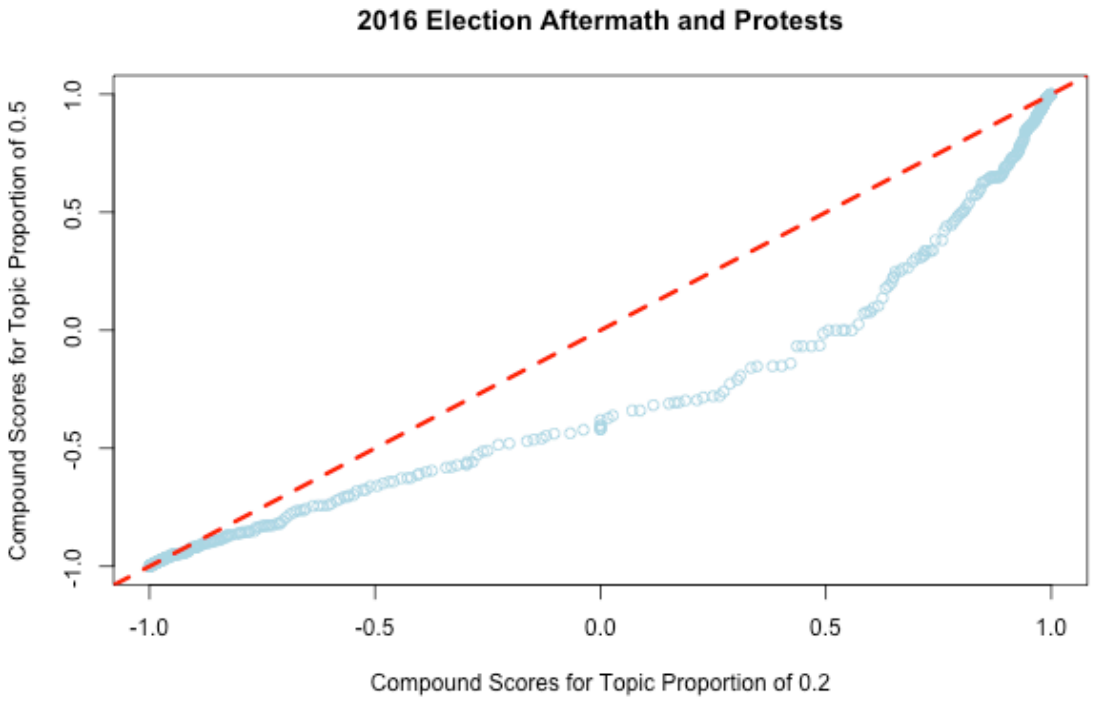
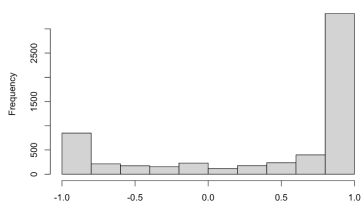


Figure 5.3: QQ plot for the Presidential Election Aftermath and Social Unrest

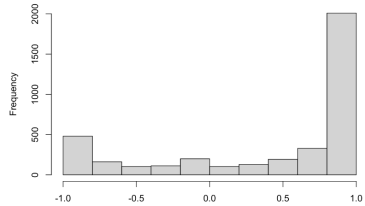
compound scores based on the threshold, we elected to go with a threshold of 0.2. A deeper investigation into an appropriate value for the threshold will be justified in Section 5.2.1. The resulting polarization plots are shown in Figure 5.4.

The polarization plots for the Presidential Election topic are very similar in shape to those for all articles during these periods. However, the overall number of articles is fewer. These plots have the same issues noticed in the polarization plots for all articles, figure 5.1, that most articles have a compound score close to 1 or -1 with a slight mode close to zero. Given that this issue is present in the aggregate to such an extreme, it would be surprising if it were not in the different topics. As examples, consider the topics "Cooking/Food" and "Crime Report" from the same period given in figures 5.5 and 5.6. Note that we only include plots for these topics from 1988, 1996, 2008, and 2020 to give a general idea of their distribution. The plots for the remaining years are similar.

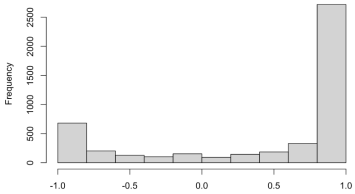
Although the polarization plots for "Cooking/Food" and "Crime Reports" do not also share the characteristic "U" shape associated with the other polarization plots that we have considered, they do still have the issue that the overwhelming majority of values are close to one extreme or the other. For the "Cooking/Food" topic, most articles have a positive sentiment and thus have compound scores close to 1, while most "Crime Report" articles have compound scores close to -1. This result gives a good "sanity check" of our process as one would expect a topic on cooking or food, in general, to be primarily positive, using positive words to describe the food or recipes while avoiding mostly negative sentiment words. On the other hand, one would expect a topic about crime to be inherently negative, given the nature of the



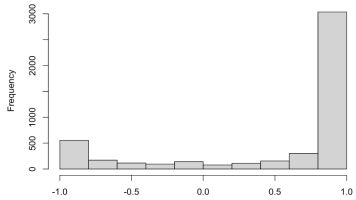
(a) Presidential Election 1988



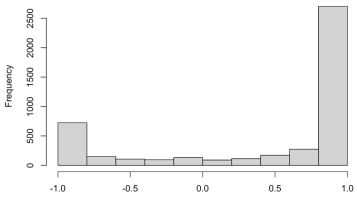
(b) Presidential Election 1992



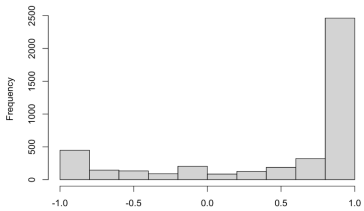
(c) Presidential Election 1996



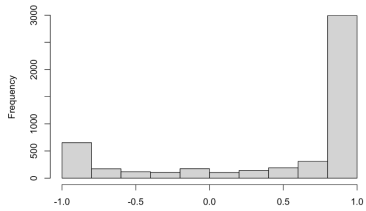
(d) Presidential Election 2000



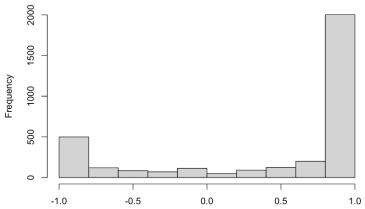
(e) Presidential Election 2004



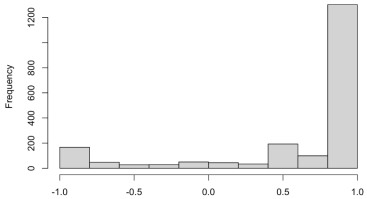
(f) Presidential Election 2008



(g) Presidential Election 2012

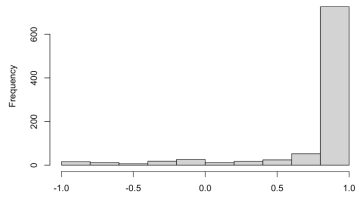


(h) Presidential Election 2016

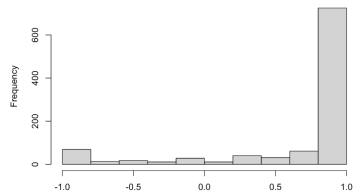


(i) Presidential Election 2020

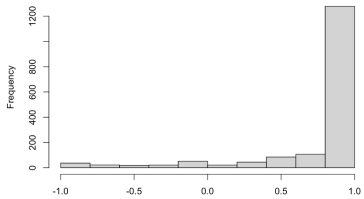
Figure 5.4: Polarization plots for all articles within the "Presidential Election" topic covering the 15 days preceding a US election, election day, and the 14 days following the election.



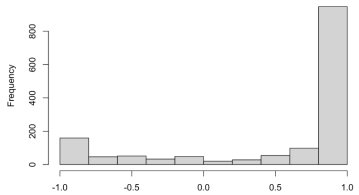
(a) Cooking/Food Topic 1988



(b) Cooking/Food Topic 1996

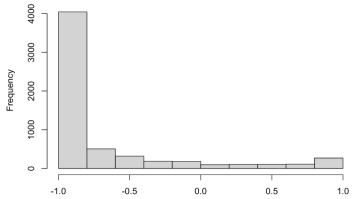


(c) Cooking/Food Topic 2008

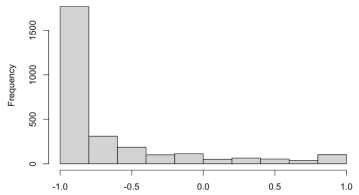


(d) Cooking Food Topic 2020

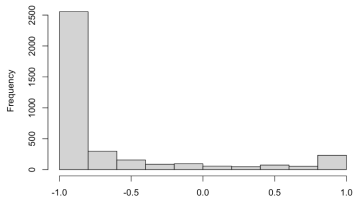
Figure 5.5: Polarization plots for all articles within the "Cooking Food" topic covering the 15 days preceding a US election, election day, and the 14 days following the election.



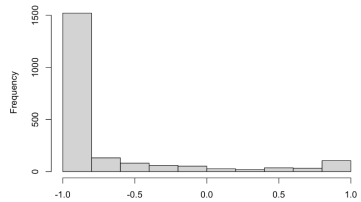
(a) Crime Report Topic 1988



(b) Crime Report Topic 1996



(c) Crime Report Topic 2008



(d) Crime Report Topic 2020

Figure 5.6: Polarization plots for all articles within the "Crime Report" topic covering the 15 days preceding a US election, election day, and the 14 days following the election.

events discussed in the article.

Initially, our goal was to use characteristics of these distributions, such as measures of center, variance, skewness, or range, to describe the polarization of the given topic; however, as we observed, most of the articles contain values close to the extremes, meaning that most of these measures do not change very much. We could increase these compound scores by changing the α value mentioned in section 4.2, but without some theoretical framework to select an appropriate value, changing this value could have the impact of changing potential conclusions and asking questions of the efficacy of our process. To avoid the influence that selecting an α value could have on the analysis, we decided to turn the compound score into a binary variable based on the sign of the compound score. This binary variable allows us to avoid problems in selecting an appropriate value for α. It is justifiable as the loss of information is negligible, since most of the compound scores are already near the extremes. We can still measure and quantify polarization according to our definition in Section 3.1 as we are interested in the disparity in discourse surrounding a topic. As illustrated in Figure 3.3, we define a topic as polarized when there is a roughly 50/50 split in positive vs. negative sentiment in the discourse surrounding a topic. By modeling polarization as the percentage of articles with positive sentiment, we can view significant changes in this measure as indicating that the topic has become more or less polarized.

In addition to analyzing polarization plots surrounding Presidential Elections, we further investigated the polarization surrounding the Summer Olympics, which generated results remarkably similar to those presented here. The polarization plots for

these events and discussion are in the Appendix B.

5.2 Quantifying Polarization: A Time Series Approach

As discussed in Section 5.1, our initial attempt to discover changes in polarization plots surrounding anticipated polarizing events, such as the Presidential Election, Summer Olympics, and 9-11, did not yield the expected results. The changes in the polarization plots from one Presidential Election to another were minimal at best. The same is true for the reoccurring topics that appear concurrently with the Presidential Election topic, such as Cooking/Food and Crime. Given these results, we decided to analyze all of the available data instead of singling out events and looking for changes in polarization. In this way, we can use the entire data set as the background from which we can make comparisons, and once fully developed, having this framework in place allows us the ability to investigate any number of events, even if we did not anticipate them as having high or low polarization.

As in event-based investigations, we still needed a way to "group" articles to fit the topic model. The grouping of articles by month of publication offered several benefits. For example, it gives a natural clustering of articles that is intuitive to most people. Additionally, the number of articles published monthly is significant, thus allowing for a stable estimation of quantities of interest. Our data set contained 119 months worth of articles, each month requiring a unique STM topic model fit

for a pre-specified number of topics. As in event-based exploration, we used $k = 20$ topics per month. See Appendix A for an analysis and justification for this number of topics. Thus, in total, there were $2,380$ topics that needed to be labeled according to the process outlined in Section 4.5.1. However, given the sheer number of topics, we needed a way to automate the matching process, as manual matching on that scale would be nearly impossible. We developed the process for matching topics, presented in Subsection 4.5.2. For this framework to be practical, we must compare $\binom{2,380}{2} = 2,831,010$ topics. Once this framework is in place, analyzing any number of topics or events will be possible and significantly more efficient than generating data and models for single events. We use this framework and modeling process to develop the following polarization analyses.

Before analyzing various topics/events for signatures of polarization, we first consider the issue of document/topic proportion under this new framework.

5.2.1 Document/Topic Proportion Threshold

A feature of topic models such as LDA and STM is that they allow for estimating the proportion of a document devoted to a topic. Although this matches intuition by allowing articles to contain multiple topics, it leaves us with the task of determining a proper threshold for which to include an article in a "topic". As discussed in Section 3.1 and illustrated in Figure 3.3, our method matches articles based on a common topic and then compares the sentiments of these articles. Thus, our threshold issue is equivalent to asking "what proportion of an article should be devoted to

a topic in order for that article to contain the topic?" This sounds much more like a philosophical question than a statistical one. However, we need to evaluate possible threshold values and determine their effect, if any, on the possible polarization of a topic.

In a sense, we are "jumping the gun" with this analysis by using our response of interest, the percentage of articles with positive sentiment, to justify our conclusions about topic/document proportions before justifying said response of interest. However, this presents a "chicken and egg" situation as we cannot evaluate different responses of interest without establishing a topic/document threshold. We are presenting the analysis for the topic/document threshold first, as this value is necessary for generating polarization plots, which all other responses require.

Given that most of the sentiment scores are 1 or -1, creating a binary variable from these sentiment scores is not an extreme idea; however, a complete analysis and comparison of different responses and why the percentage of articles with positive sentiment was selected are presented in Section 5.2.2.

By threshold, we mean the minimum topic proportion a document/article must obtain for inclusion in the topic analysis. In order to examine the effect different proportions can have on our response of interest and the percentage of articles with a positive sentiment, we selected twenty random topics. We computed the percentage of articles with positive sentiment for the threshold values of $.1, .11, .12, \ldots, .59, .6$. The results are illustrated in Figures 5.7 and 5.8, where Figure 5.7 shows the topics in chronological order and Figure 5.8 shows the topics ordered by decreasing positive

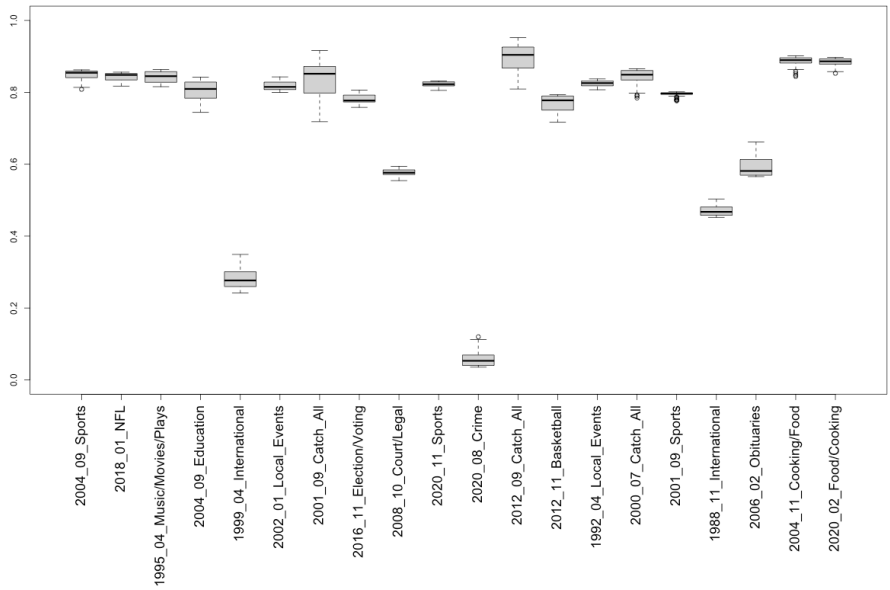


Figure 5.7: Box plot of percent positive sentiment for 20 randomly selected topics.

sentiment.

Figure 5.8 illustrates that the percent positive sentiment for the randomly selected topics matches our expectations. For example, the discussion around most topics uses positive sentiment, such as sports, cooking/food, and local events/announcements. At the same time, the remaining topics would probably have more negative sentiment associated with them, such as obituaries, international news, and crime. Figure 5.7 shows that the variation within a topic's percent positive sentiment is slight compared to the overall variation between topics. This difference in variation is essential to note, as it indicates that the observed changes in the percent of positive sentiment over time are not due to the choice of threshold for document/topic proportion. However, a few topics have a non-negligible spread over the range of document/topic

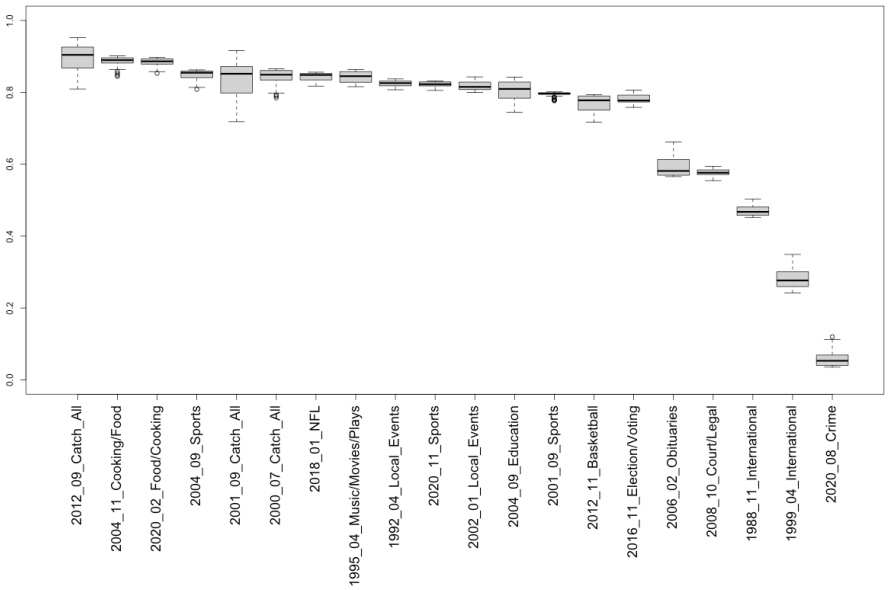


Figure 5.8: Box plot of percent positive sentiment for 20 randomly selected topics ordered by median percent positive sentiment.

proportion used.

The topics with the most extensive spread, Education (09/2004), International News (04/1999), Catch All (09/2001) (this is a topic of words common to all articles, such as, ”said,” ”they,” ”think”, etc.), Crime (08/2020), and Obituaries (02/2006) are plotted in Figure 5.9. This figure shows how the percent positive sentiment changes based on the document/topic proportion threshold. Ideally, this plot would be “flat” over some interval of “reasonable” document/topic proportions, such as 0.2 to 0.5. While these plots are not “flat” over any such interval, the differences between the percent positive sentiment when the threshold is 0.2 versus when the threshold is 0.5 is small relative to the difference in percent positive sentiment between the topics. Note Figure 5.9 shows a slight “bend” in each graph as the document/topic propor-

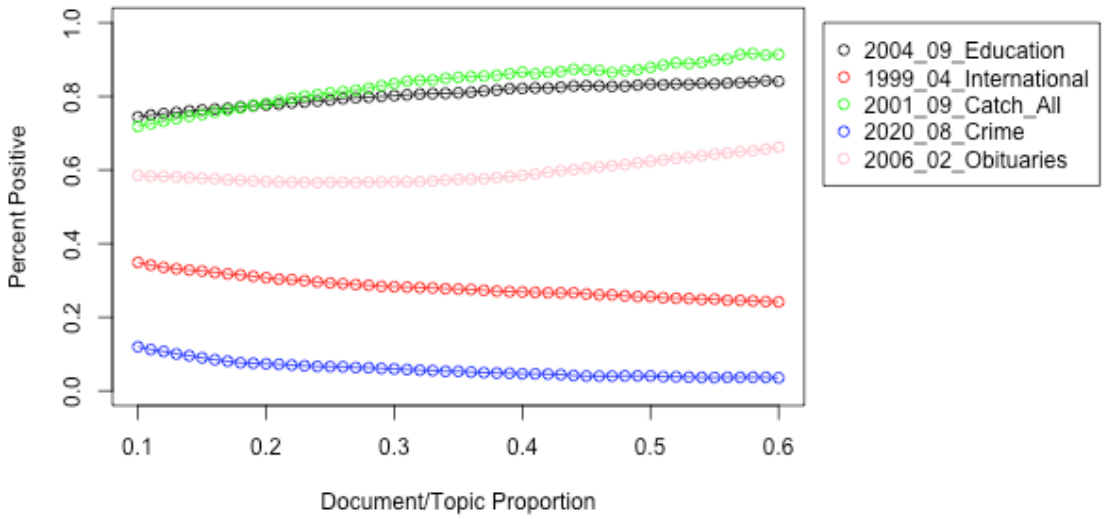


Figure 5.9: A scatter plot of the changes in percent positive sentiment as document/topic proportion threshold changes from 0.1 to 0.6

tion threshold decreases. This bend happens because as we decrease the threshold, more articles are included in the calculation of percent positive sentiment. Thus, this metric approaches the overall percent positive sentiment for all articles included in that month's topic modeling.

As mentioned in Section 4.1, we selected articles for analysis based on anticipated periods of high and low social polarization. Given this selection, we deemed it necessary to explore the effect the threshold value of document/topic proportion has on the percent positive sentiment for these selected topics. In total, we handpicked 23 topics for examination. The percentage positive sentiment distributions for these topics are presented in Figures 5.10 and 5.11. As before, the variation within a topic is relatively small compared to the variation between topics. Also, like before, some topics have a noticeablely extensive range of values.

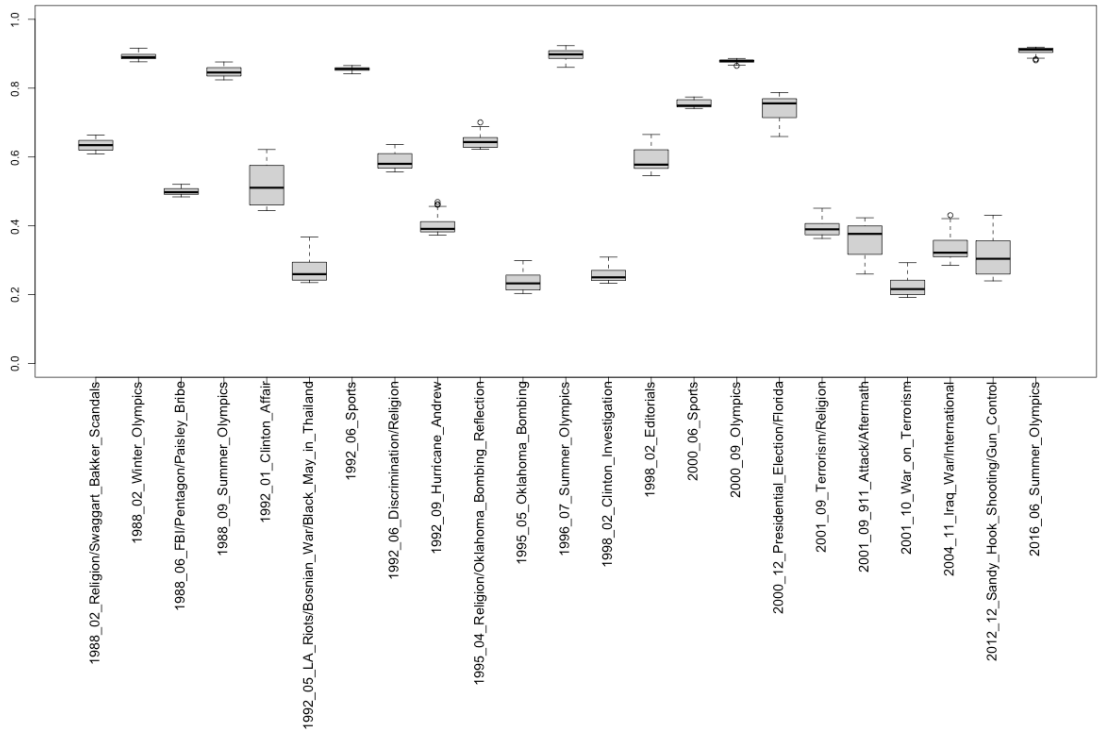


Figure 5.10: Box plot of percent positive sentiment for 23 handpicked topics.

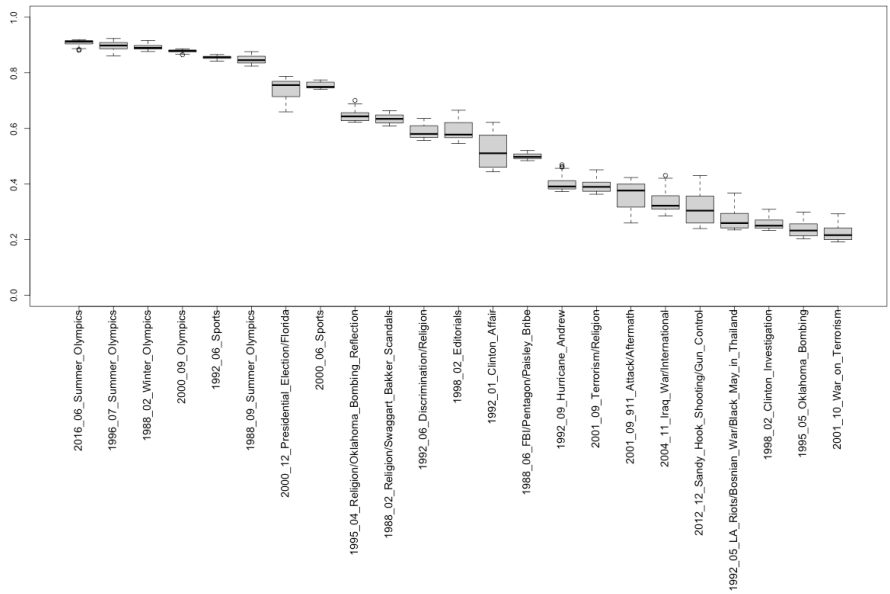


Figure 5.11: Box plot of percent positive sentiment for 23 handpicked topics ordered by median percent positive sentiment.

The topics with the highest spread are Bill Clinton affair allegations (01/19992), the topic of the LA Riots / Bosnian War / Black May in Thailand (05/1992), Editorials (02/1998), the Presidential Election and Florida Recount decision (12/2000), the Aftermath from the 9/11 attacks (09/2001), and the Sandy Hook School Shooting (12/2012). These topics are explored in more detail in Figure 5.12. For most topics, the scatter plot resembles the plots in Figure 5.9. However, there is one that stands out: the Bill Clinton affair allegations topic from January of 1992. There is a noticeable increase in the percentage of positive sentiment for this topic as the document/topic proportion approaches 0.6. Upon investigating this phenomenon, we discovered that the number of articles included in this topic based on the increasing threshold has decreased substantially. As a point of comparison, the Clinton affair

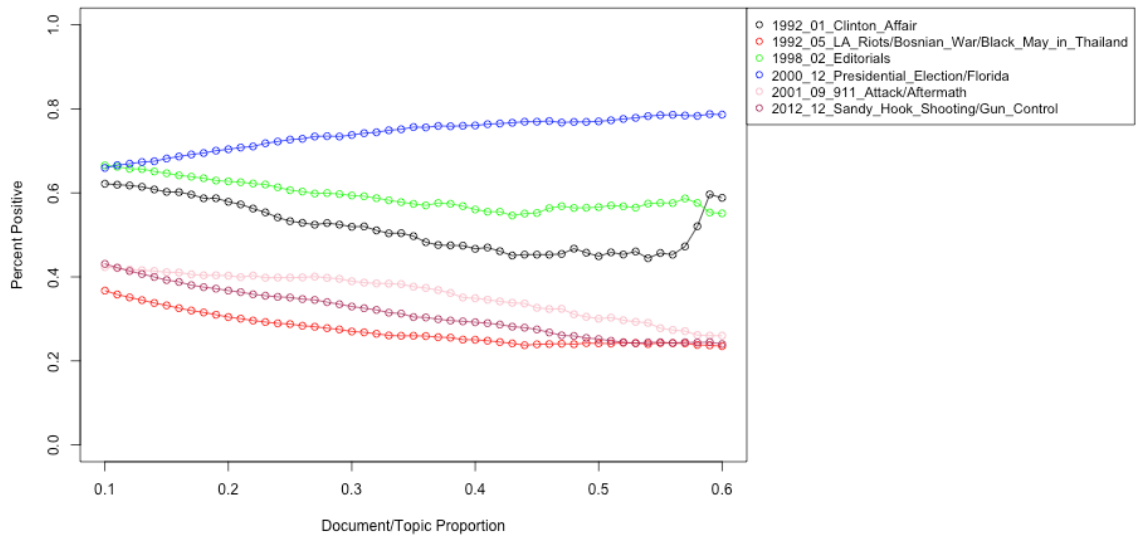
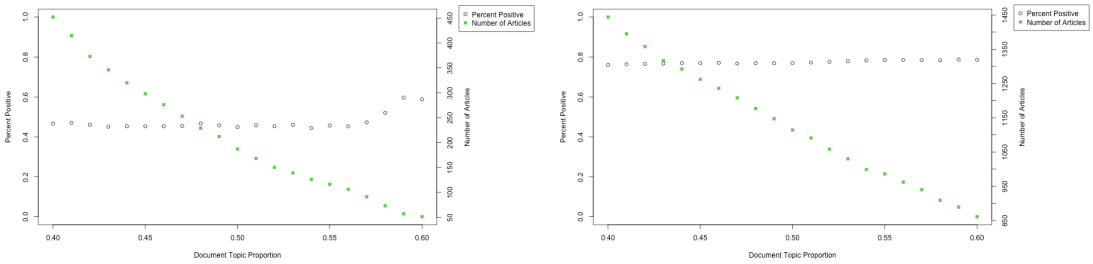


Figure 5.12: A scatter plot of the changes in percent positive sentiment as document/topic proportion threshold changes from 0.1 to 0.6

allegations topic from January 1992 and the Presidential Election topic from December 2000 are shown in Figure 5.13. For reference, we plotted the percent positive sentiment on the left axis and the number of articles included in the topic on the right axis. The number of articles on the Clinton affair allegations topic is considerably lower than those on presidential elections. This explains the sudden change in the percentage of positive sentiment for this topic.

As mentioned earlier, most topics examined in this section show a trend toward the percent positive sentiment of all articles included in the topic modeling as the document/topic proportion threshold decreases. The general trend is that as the document/topic proportion increases, the percentage of positive sentiment for topics becomes more extreme. That is, "positive" topics become more positive, and "negative" topics become more negative, in terms of the percent positive sentiment



(a) The percent positive sentiment and number of articles included based on document/topic proportion for the Clinton affair allegations topic in January of 1992.

(b) The percent positive sentiment and number of articles included based on document/topic proportion for the Presidential Election/Florida recount topic in December of 2000.

Figure 5.13

.

metric. This feature accentuates the differences between topics as we increase the document/topic proportion threshold. Thus, a conservative approach to detecting polarization changes in a topic over time or region would suggest using a smaller value for the document/topic threshold, as this would reduce the type-one error rate compared to an analysis that uses a larger value for the document/topic threshold. We shall use a threshold for document/topic proportion of 0.2 for the remainder of the results presented unless otherwise stated.

5.2.2 Metric Comparison: Range, Variance, Skewness, Percent Positive, Entropy, Gini

In January of 1998, several news outlets reported that then-President Bill Clinton engaged in a sexual relationship with a White House intern, Monica Lewinsky, between the years of 1995 and 1997. President Clinton almost immediately denied that

such a relationship existed, as one can imagine a story of this magnitude stayed in the news cycle for several months. We have data for January, February, and March that year because we selected these months for the Winter Olympics in February. Each of these months contains a topic that is at least in part about the alleged affair between Clinton and Lewinsky. We use the topic labeled "Clinton Investigation" from February of 1998 to illustrate similarities and differences between the proposed metrics to quantify polarization based on the polarization plots for each topic.

The metrics considered are the range, variance, skewness, and percent positive of the polarization distributions generated for each topic. Additionally, we consider entropy and Gini, which are transformations of the percent positive metric. As a way to introduce the time series for these topics and as a way to compare these metrics, we present the following Figures 5.14, 5.15, 5.16, and 5.17. Each plot was constructed by including all topics that matched the Clinton Investigation topic based on the method of Section 4.5.2. Each metric is computed directly from the polarization distributions for the corresponding topics. We have excluded the full names of the topics that matched in the plot for ease of visualization. Table 5.2 presents a table that contains all the labels of the matched topic. For additional time-series plots, we shall only include the unique labels that appear unless otherwise needed to emphasize. The dashed vertical lines in each plot represent a break in the time series. That is, the data jumps more than one month in time. These breaks could be due to a lack of Newbank data or topics that do not match for a given month.

Table 5.2: List of topic labels that the Clinton Investigation topic from February of 1998.

1988 01 Crime	1988 03 Crime
1988 06 Court/Legal	1988 07 Court/Legal
1988 08 Court/Legal	1988 09 Court/Legal
1988 10 Crime/Court/Legal	1988 11 Court/Legal
1988 12 Court/Legal	1992 01 Crime
1992 02 Crime	1992 04 Court/Legal
1992 05 Crime	1992 06 Court/Legal
1992 08 Crime	1992 09 Crime
1992 10 Court/Legal	1992 11 Court/Legal
1994 01 Court/Legal	1994 02 Court/Legal
1995 04 Court/Legal	1995 05 Crime
1996 06 Court/Legal	1996 07 Court/Legal
1996 08 Crime	1996 09 Crime
1996 10 Court/Legal	1996 11 Crime
1996 12 Crime	1998 01 US Politics/Clinton Investigation
1998 02 Clinton Investigation	1998 03 Crime/Clinton Investigation
1999 03 Crime 1999 04 Crime	2000 06 Crime
2000 07 Crime	2000 08 Crime
2000 09 Crime	2000 10 Crime 2000 11 Crime
2001 08 Crime	2001 10 Crime

2002 01 Court/Legal/Enron	2002 02 Court/Legal
2002 03 Crime	2004 06 Court/Legal
2004 07 Court/Legal	2004 08 Crime
2004 09 Court/Legal	2004 10 Court/Legal
2004 11 Court/Legal	2006 01 Court/Legal
2006 02 Crime	2006 03 Court/Legal
2007 03 Crime	2007 04 Crime
2007 05 Court/Legal	2008 08 Court/Legal
2008 09 Court/Legal	2008 11 Crime
2008 12 Court/Legal	2012 06 Crime
2012 10 Court/Legal	2012 11 Crime
2012 12 Crime	2013 07 Crime
2013 08 Court/Legal	2016 06 Legal/Courts
2016 07 Legal/Courts	2016 08 Legal/Courts
2016 09 Legal/Courts	2016 10 Crime
2016 11 Legal/Courts	2018 01 Crime
2018 02 Court/Legal	2018 03 Court/Legal
2018 04 Court/Legal	2020 01 Legal/Courts
2020 02 Crime	2020 03 Crime
2020 06 Legal/Courts	2020 08 Legal/Courts
2020 09 Legal/Courts	2020 10 Legal/Court
2021 01 Legal/Courts	2021 05 Legal/Courts

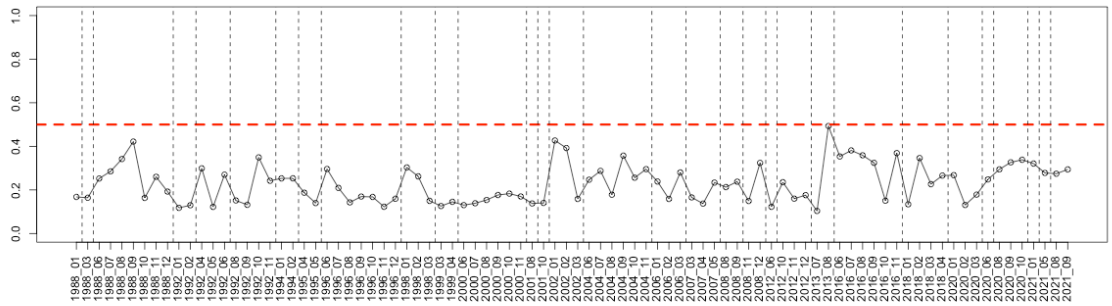


Figure 5.14: Time Series of the percent of documents that have positive sentiment for all topics which match the Clinton Investigation topic from February of 1998.

2021 08 Legal/Courts	2021 09 Legal/Courts

There are several observations to draw from these plots and the table of topics. The first is that the overwhelming majority of topics that matched the Clinton Investigation topic are either crime or court/legal related. This matches intuition, as one would expect these topics to overlap in frequently used words. The difference between these Crime/Court topics and the Clinton Investigation topic is the unusual nature of having a sitting President under investigation. When comparing potential metrics, the obvious observation is that the range is a useless metric to consider in the context of polarization, as it is constant for all of the 88 topics included in the graph. This is consistent with our observations from the polarization plots observed in Section 5.1. Thus, we shall no longer consider the range as a possible metric for quantifying polarization.

The second observation is that there is little information readily apparent from the

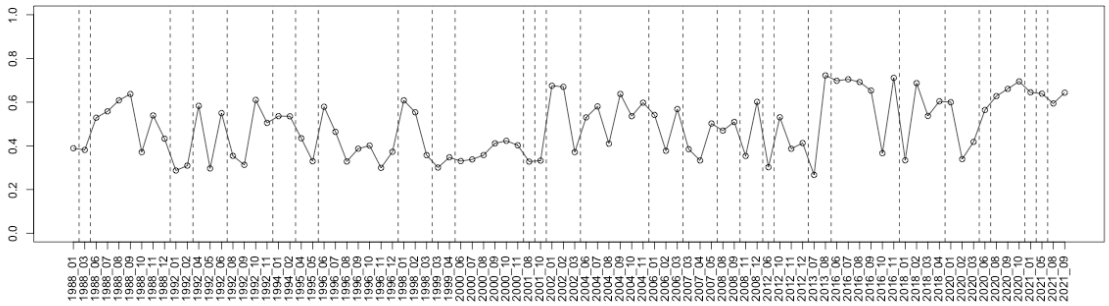


Figure 5.15: Time Series of the variance of the polarization distributions for all topics which match the Clinton Investigation topic from February of 1998.

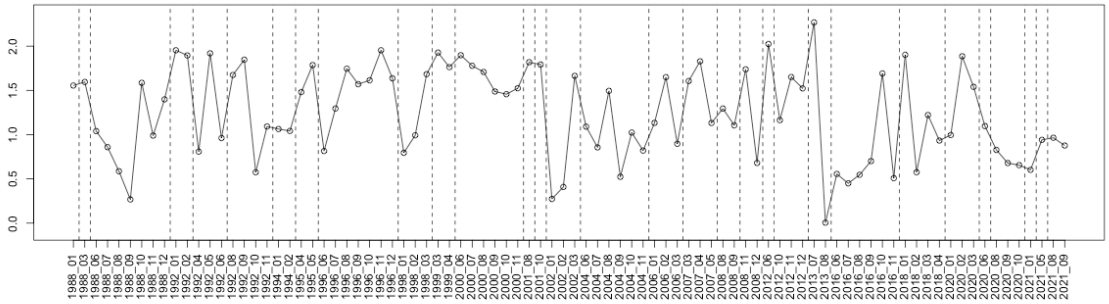


Figure 5.16: Time Series of the skewness of the polarization distributions for all topics which match the Clinton Investigation topic from February of 1998.

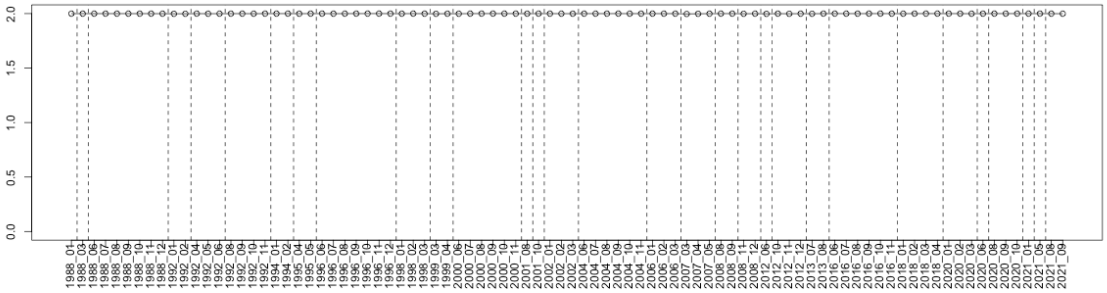


Figure 5.17: Time Series of the range of the polarization distributions for all topics which match the Clinton Investigation topic from February of 1998.

time series plot on skewness. The first issue is that the scale for skewness is unintuitive. It is unclear what it means in the context of polarization for the skewness to change from greater than 2 to essentially 0 in a month, as occurs from July of 2013 to August of the same year. The problem is that the skewness does not fit our definition of polarization in Section 3.1. For example, a polarization distribution can have the same number of 1's as -1's and no other values. In this case, the skewness of this distribution is 0. However, it is possible to have the same skewness value for a distribution split evenly between 0.5 and 1. However, according to our definition of polarization, the first hypothetical distribution would be highly polarized. In contrast, the second distribution would be significantly less polarized, as there is general agreement on the sentiment used to discuss the topic. Given these shortcomings, skewness is not an appropriate metric for quantifying polarization.

The third observation is how remarkably similar the variance and percent positive plots are. Both plots tend to increase and decrease simultaneously and in similar

ways, and pass the eye test to emphasize similar trends in the data. Also, unlike skewness, both measures have a more interpretable scale, although percent positive is easier to understand. In terms of scale, we know that both metrics are between 0 and 1 for the data under consideration. Percentage naturally obtains these bounds, and the variance is always bounded below by zero. The upper bound on variance comes from Popoviciu's inequality on variances, which applies to any bounded probability distribution. Popoviciu's inequality is $\sigma^2 \leq \frac{1}{4}(M - m)^2$ where M is the upper bound on the distribution and m is the lower bound on the distribution, see [Pop35]. In our constructed polarization distributions, $M = 1$ and $m = -1$. Thus, $\sigma^2 \leq \frac{1}{4}(1 + 1)^2 = 1$. However, one thing differentiates the percentage of documents with positive sentiment and the variance of the polarization distributions. That is, the differentiation between positive topics that become more negative and negative topics that become more positive. In order to better understand this difference, we shall consider another topic and its resulting time-series plots.

For comparison, let us consider the US Politics topic from the same period, March of 1998, and compare the percent positive and variance time-series plots for this topic to the ones based on the Clinton Investigation topic. These plots are shown in Figures 5.18 and 5.19.

Notice that the topics that match the US Politics topic from March 1998 are overall positive, with a percent positive sentiment greater than 0.5 for most topics. There is a perceptible drop in May 1999, when the metric dips below 0.5. These plots are significantly less similar to the percent positive and variance plots for the Clinton Investigation time series.

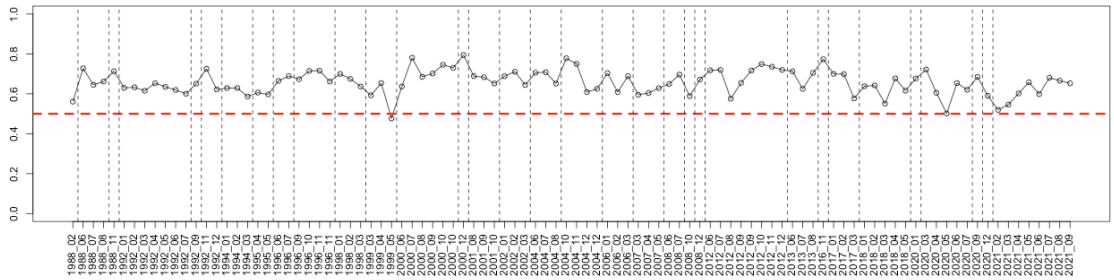


Figure 5.18: Time Series of the percent of documents that have positive sentiment for all topics which match the US Politics topic from March of 1998.

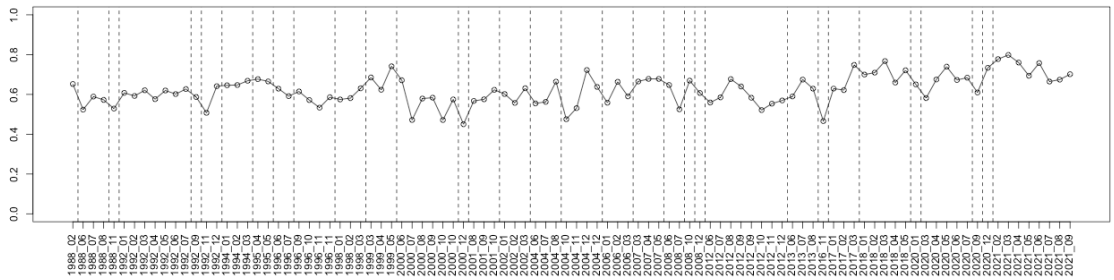


Figure 5.19: Time Series of the variance of the polarization distributions for all topics which match the US Politics topic from March of 1998.

In this example, the relationship appears to be inverted; as the percent positive decreases towards 0.5, the variance increases and vice versa. This inversion fits our expectations, given that most sentiment values are either 1 or -1. The decrease in the percent positive value towards 0.5 is caused by increasing the number of -1's in the distribution, thus increasing the variance. Initially, this seems like a benefit for using the variance metric to compare topics, as it appears to match our intuition in that as polarization, defined as a disparity in sentiment use, increases, the variance also increases. However, there is a loss of directional information when using the variance metric.

Consider Figures 5.15 and 5.19. Both time series obtain very similar values and have peaks that could be indications of polarization, but the similarities overlook that both time series have very different "background" sentiments. Overall, the time series for the Clinton investigation topic has a negative sentiment, and the topic of US politics has a positive sentiment. Knowing these background sentiments matters in determining the degree of change in polarization. Consider the following hypothetical as an illustration of this idea. Consider a generally negative topic, such as the Clinton Investigation, and the crime/legal topics that match it. Suppose that the next unseen entry in this time series obtains a percent positive value of 0.47. According to our definition of polarization, this would indicate that this topic has become more polarized compared to the topic's baseline sentiment. Now consider that this same entry instead obtains a percent positive of 0.53. This value would be the same "distance" from our 50-50 boundary. However, a percent positive of 0.53 is a much more dramatic change for a generally negative topic than is a value of 0.47.

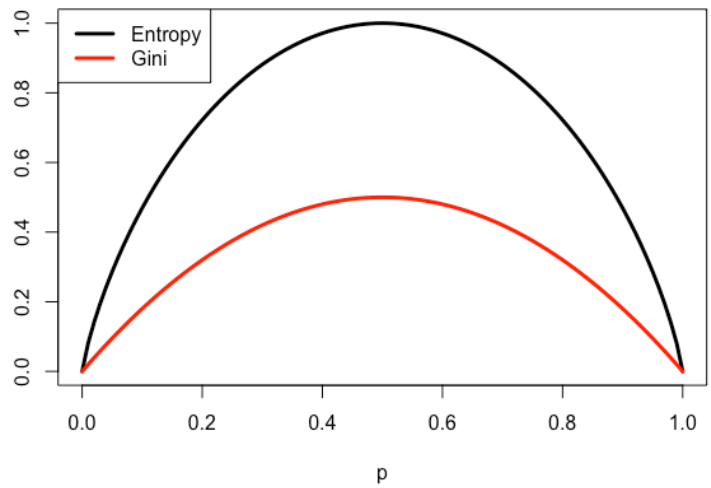


Figure 5.20: The Entropy and Gini functions for a random binary variable.

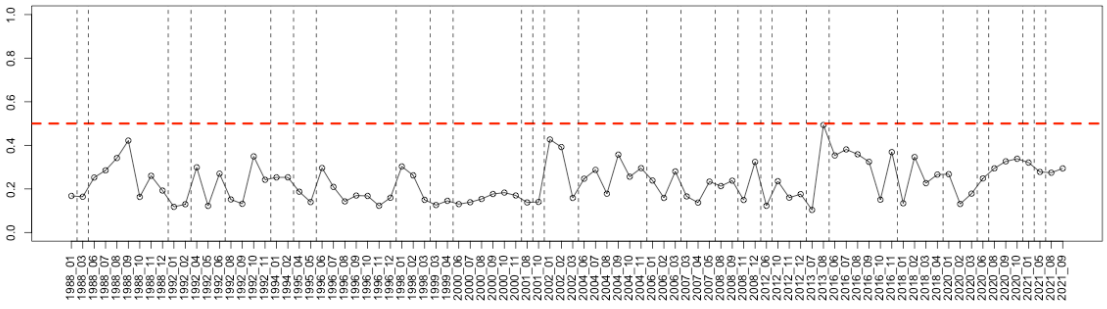
Probably, the variance of the polarization distribution with a percent positive of 0.47 is similar to that of the polarization distribution with a percent positive of 0.53. Even if the variances are not similar, a change in a topic from a generally negative sentiment to a positive one (or vice versa) is lost if one uses the variance metric. Thus, given that the percent positive metric contains the same information as the variance and this additional information when the majority articles swap sentiment, we have decided to use the percent of positive documents as our preferred metric.

As mentioned earlier, we considered two other metrics, entropy and Gini, both of which are a function of the percent positive metric. For a binary random variable, X, the entropy associated with that random variable is given by $H(X) = -p\log_2(p) - (1-p)\log_2(1-p)$, where p is the probability of the "success" outcome of the binary variable. Given this definition, the entropy is bounded between 0 and 1, with 0

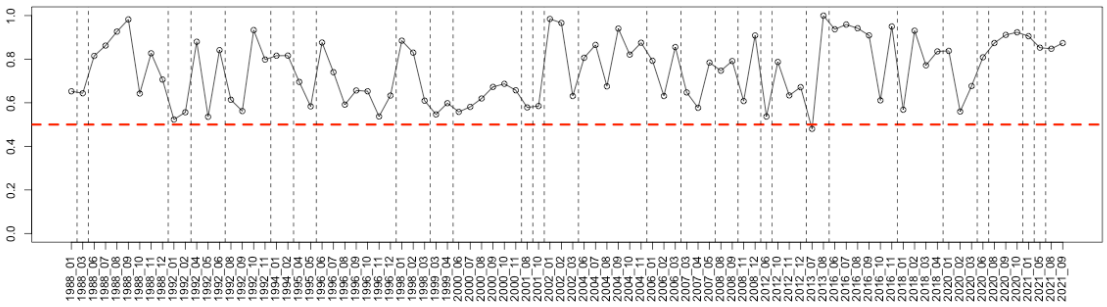
occurring when $p = 0$ and $p = 1$ and an entropy of 1 when $p = 0.5$. Gini has a similar definition for a binary variable given by $G(X) = 1 - p^2 - (1 - p)^2$. Gini is bounded between 0 and 0.5, occurring at $p = 0, 1$ and $p = 0.5$, respectively. Both functions are plotted in Figure 5.20 as a reference point.

For illustrative purposes, Figure 5.21 presents the Clinton investigation topic time series with the three metrics: positive sentiment, entropy, and Gini. As expected, the entropy and Gini time series are extremely similar to the percent positive time series, which is unsurprising given that both are direct functions of the percent positive metric. From the formulas for entropy and Gini and the time-series plots presented, it is clear that these metrics suffer from the same issue as the variance metric, possibly to a greater extent. That is, both metrics are centered on 0.5 and do not offer a way to distinguish a topic with a percentage positive of 0.45 from a topic with a percent positive of 0.55. This lack of distinction becomes important when using all the months that match a specific topic to get a "baseline" for the typical sentiment a topic carries. A typically negative sentiment topic containing an entry of 0.55 in the percent positive value is a much more drastic change than 0.45. However, both are "equivalent" when using the entropy or Gini metrics.

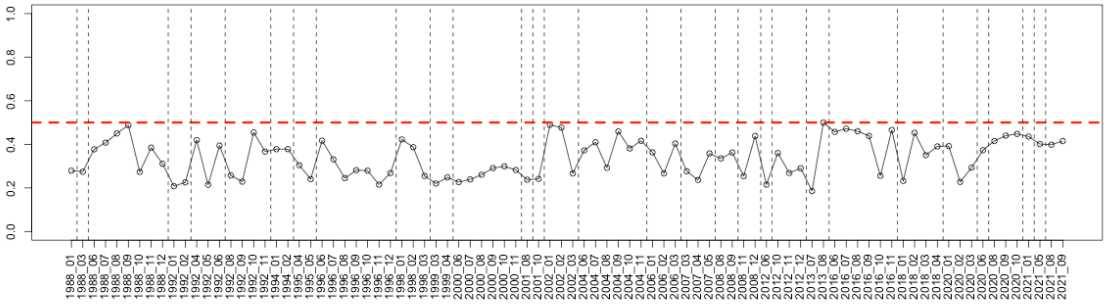
Numerous metrics could be used to quantify polarization; however, of all the ones considered, the percentage of articles with positive sentiment is the most consistent and provides the most information. For this reason, we shall be using this metric exclusively to evaluate polarization.



(a) Percent Positive



(b) Entropy



(c) Gini

Figure 5.21: The Clinton Investigation topic time-series for percent of articles with positive sentiment, entropy, and Gini.

5.3 Examining Time Series Plots for Signatures of Polarization

With the time-series framework now established, we present evidence of polarization captured in time series plots. We could construct 2380 unique time series plots, one for each topic. This is obviously way more plots than is feasible to include in this document. The examples chosen were selected on the basis of demonstration purposes and the author's curiosity.

5.3.1 US Politics January 2020: First President Trump Impeachment

Political topics in the US are obviously of interest when analyzing polarization. The coverage of US political news is extensive, as each month includes thousands of articles, and the discussion around this topic can change rapidly based on current events. For example, we selected the "US Politics/Trump Impeachment" topic from January 2020. During this month, much of the discussion on this topic centered around the traditional expected discourse in US Politics. For example, several articles discuss gun control at the regional and national level. There are other articles on civil rights, specifically those related to abortion and LGBTQ issues. However, this month stands out as on January 16, 2020, articles of impeachment were sent from the House of Representatives to the US Senate, charging then President Trump with abuse of power and obstruction of Congress related to his taped phone call to

Ukrainian President Volodymyr Zelensky. Given that President Trump is only the third president in US history to be impeached, we wanted to investigate this topic and examine how it compares to the other topics from the NewsBank data set.

The time series for this topic and all the matched topics are shown in Figure 5.22. The majority of these topics have been labeled "US Politics," with a few having additional descriptors included, such as "Marriage Equality," "Elections Fraud Claims", "Trump Impeachment," "Absentee Voting/Turnout," "Trump Second Impeachment/January 6". The remaining topics have labels that are directly related to "US Politics", such as, "Election/Voting," "Local Elections," "Tax/Budget/Legislation," and "Gun Legislation." The "US Politics/Trump Impeachment" topic from January 2020 has a similar percentage positive value to most of the other matched topics, with a percentage positive value over 0.6. Several factors could explain why this topic, even with such an anomalous event contained within, has a typical value of percent positive compared to other "US Politics" topics. First, the news that led to the impeachment broke in September 2019, four months before the debate being considered. Four months is a considerable amount of time for the authors of these articles to discuss and examine the implications of this discovery. A second possible explanation is that the House of Representatives voted to impeach President Trump on December 15, 2019, a whole month before the articles of impeachment were delivered to the Senate and the trial could begin. Again, they are giving the authors plenty of time to discuss the implications and disagreements present on this topic. Unfortunately, we do not have access to data from September or December 2019 and therefore cannot verify these hypotheses.

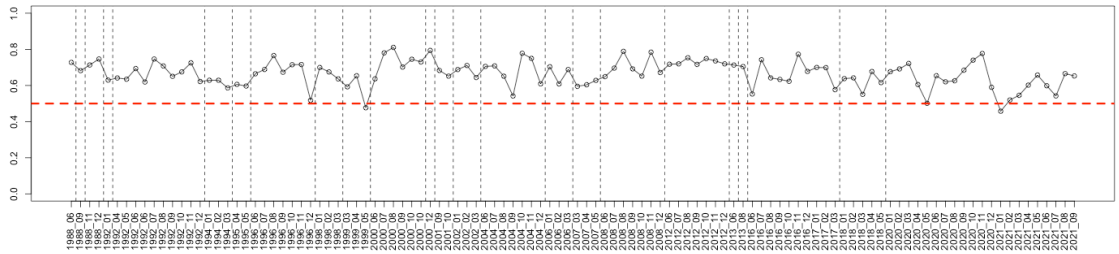


Figure 5.22: Time series for the US Politics/Trump Impeachment topic from January 2020 and all the topics that match.

Even with the main topic, "US Politics/Trump Impeachment," from January 2020, having a larger than expected percent positive value, there are several notable deviations from the norm for this topic that are worth exploring. These topics are relatively positive, with most topics above the percent positive value of 0.6, as demonstrated by the histogram of the percent positive values for these topics 110 shown in Figure 5.23. However, there are a few discernible exceptions in which the time series dips close to a percentage positive value of 0.5 and a couple of times where it drops below 0.5. In order to demonstrate the effectiveness of our methodology, we shall highlight these months and discuss potential explanations for the observed polarization.

The months of May 1999 and June 2016 both show significant dips in sentiment, and we hypothesize that both dips are due to similar changes in the discourse. On April 20, 1999, two high school seniors committed what was, at that point, the deadliest school shooting in US history at Columbine High School. This massacre shocked the nation and subsequently became a point of contention as the motives and background of the shooters came to light. At this time, gun control and regulation by the Federal

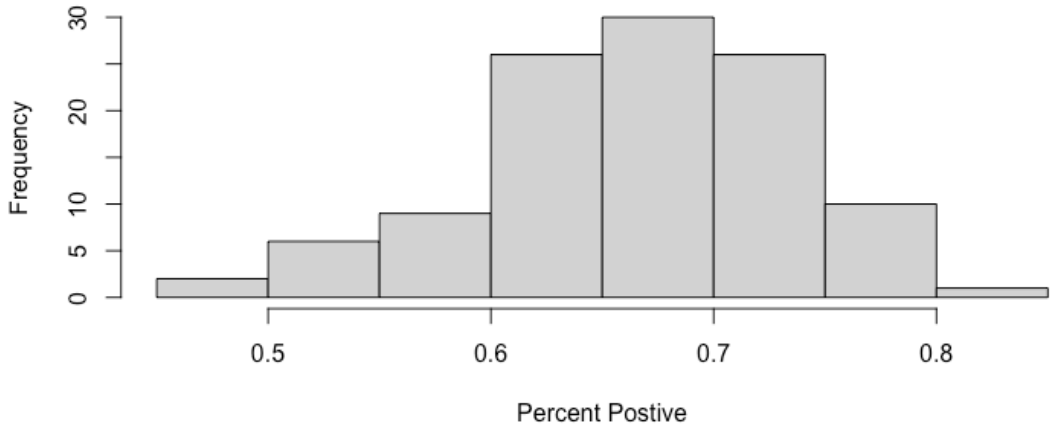


Figure 5.23: Distribution of Percent Positive values for the 110 topics matching the "US Politics and First Trump Impeachment" from January of 2020.

government had been part of the political debate for nearly a decade, starting with the Crime Control Act of 1990, followed by the Brady Handgun Violence Prevention Act passed in 1994. However, the events at Columbine High School pushed this debate to the forefront of American discourse. We illustrate this push by highlighting some of the headlines for the "US Politics" topic in May of 1999:

- "Gun laws don't stop crime"
- "Gun rights apply to citizens"
- "Gun Control Battle Drawing Up Sides in Congress, Society"
- "Gun control becomes a political bull's-eye. School shootings prod both parties to act"

- "Clinton wants gun law 'before school lets out'"
- "GUN-CONTROL VOTES CLOSELY LINKED TO NRA MONEY"
- "LIBERALS' HATRED OF GUNS SHOULDN'T SWAY CONGRESS"

This list represents just a tiny sample of the dozens of headlines we could have presented that directly refer to gun control. On 12 June 2016, seventeen years after the events at Columbine High School, another mass shooting occurred at the Pulse Nightclub in Orlando, Florida. Although there were several mass shootings in the years between the Columbine massacre and the massacre at the Pulse Nightclub, the issue of gun control is still front and center in 2016, as evidenced by the following headlines, which are in the "Presidential Election/US Politics" topic from June 2016.

- "Gun-control group backs Clinton: Will it make a difference in 2016 race?"
- "Gun control not the issue"
- "Democrats push for gun-control legislation in wake of Orlando attack"
- " Orlando shooting intensifies gun debate"
- " Congress stalemated on guns despite shooting, filibuster"
- "Durbin, Kirk for stricter gun measures which failed in Senate"

Again, this is only a sample of the many headlines directly relating to gun control included in this topic. Of course, gun control is not the only contributing factor to

the increased polarization present in these months, but it clearly had an impact on the increased contentiousness found in the discourse on these topics.

The May 2020 topic shown in Figure 5.22 was labeled "US Politics" and occurred two months into the nationwide lockdown for the COVID-19 pandemic. During this month, protests began to form, asking mayors, governors, and other representatives to reopen many businesses that government officials ordered to suspend operations. People's attitudes towards these lockdowns, which consisted mainly of keeping them in place or removing them, were highly correlated with political affiliation; see [Sat+22] as an example. Thus, it is consistent with our finding that this topic became more polarized during this period, since there was general disagreement on the appropriate course of action given the COVID-19 pandemic.

On 6 January 2021, supporters of President Trump stormed the US Capital building in an attempt to stop or delay the certification of the Electoral College votes, which would formalize the election of Joe Biden as president. Much has been written about the events of this day, the overwhelming majority of which is beyond the scope of this work; however, this was a very divisive moment that resulted in violence and the deaths of nine people. Thus, it is unsurprising that this topic has the lowest percent positive value of any of the topics included in the time series plot.

As a point of comparison and to delve into the intricate features of the topic matching procedure outlined in Section 4.5.2, we present the time series for the topic labeled "US Politics/Trump Second Impeachment/Jan 6th" from January of 2021. The first observation is how many fewer topics matched the "US Politics/Trump Second Im-

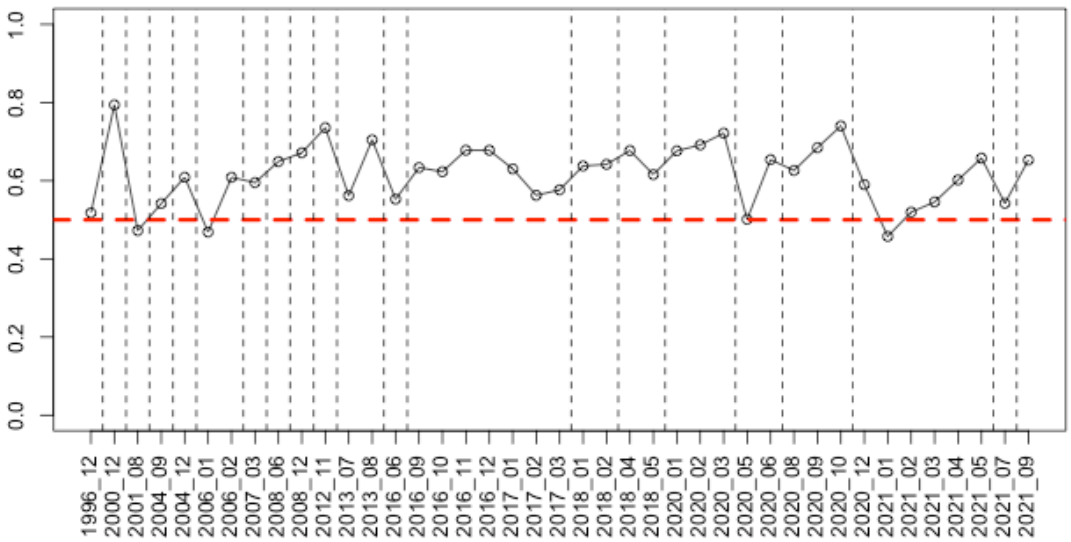


Figure 5.24: Time Series of percent positive values for each topic that matched to the "US Politics/Trump Second Impeachment/Jan 6th" from January of 2021.

peachment/Jan 6th" topic compared to the "US Politics/Trump Impeachment" topic from January 2020. As discussed in Section 4.5.2, this is due to our topic-matching procedure not having the transitive property and thus not forming an equivalence relation among topics. Both topics will appear in the time series generated by the other topic. However, there is no guarantee that additional topics that match the January 2020 topic will also match the January 2021 topic, and vice versa. There are several months contained in both time series, for example, December of 1996 and June of 2016; however, there are more "gaps" in the time series for the January 2021 topic, as this topic is an outlier in terms of the words used in the discourse around the events of the month.

5.3.2 Sports

While topics related to US Politics have a slightly positive sentiment most of the time, there is still enough disagreement that polarizing events, such as gun control debates or the discussion over reopening during the COVID-19 lockdowns, can push this topic into a highly polarized state. We have seen several examples of this in Section 5.3.1. However, other topics do not have the same baseline polarization as US politics; for example, sports-related topics have a much higher percentage of positive sentiment at baseline. Thus, polarization in this topic presents differently than in the US Politics topic.

During the 2016 NFL preseason, reporter Steve Wyche observed that Colin Kaepernick, the Quarterback for the San Francisco 49ers, was sitting while "The Star Spangled Banner" was played before the game. Traditionally, players, coaches, and staff stand during the song. When asked about why he sat during "The Star Spangled Banner" played, Kaepernick said "I am not going to stand up to show pride in a flag for a country that oppresses black people and people of color. To me, this is bigger than football, and it would be selfish on my part to look the other way. There are bodies in the street and people getting paid leave and getting away with murder." The following week, during the 49ers final preseason game, Kaepernick knelt during the US national anthem and continued to do so for the remaining 49ers games during the 2016 season. Kaepernick kneeling during the national anthem sparked a national debate on appropriate forms/ways to protest and brought national attention to the then-recent police shootings of Alton Sterling, Philando Castile, and Charles Kinsey.

The kneeling of Kaepernick during the national anthem was one of the most polarizing events to occur in sports over the last few decades. Thus, it is an excellent time to begin our investigation of polarization in sports.

To begin this investigation, we will examine the topic from September 2016 labeled "NFL/Kaepernick" and the resulting time series of matching topics. The first preseason game where Kaepernick sat during the national anthem occurred on August 26, 2016, making September the first full month of potential news articles on this topic. We present the time series for the "NFL/Kaepernick" topic from September 2016 in Figure 5.25. Note that only 21 topics are in this time series plot, which is unexpected given that most NFL topics match more than 100 other NFL and sports-related topics. As a point of comparison, the "NFL" topic from August 2016 matched 110 other topics.

The lack of matching topics is evidence that the discussion around the NFL during this period is different from the discussion around the NFL during most of the time that we have access to news articles. The baseline percent positive for this topic is significantly higher than for the US Politics topics discussed in Section 5.3.1, with most values near 0.8. The month in question has a slightly lower percent positive compared to the other months with a value of 0.75, but this is still within a reasonable range given the remaining values of the time series. This lack of disparity leaves us with the question of why signal of polarization present in our data was not stronger. Kaepernick kneeling was a highly publicized event/topic and is seemingly polarizing. We further investigate this topic based analysis to better understand sources of polarization.

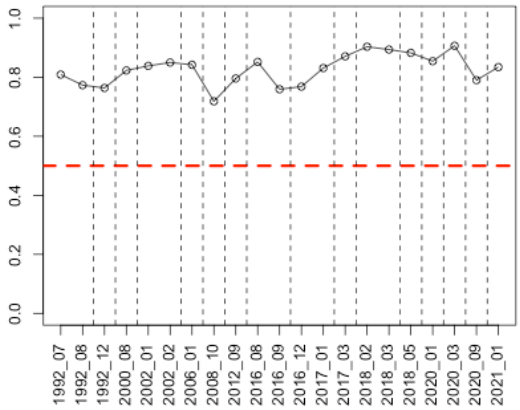


Figure 5.25: Time series for the "NFL/Kaepernick" topic from September 2016.

We searched the headlines for all articles that month and any article that contained "Kaepernick" in the headline was retained. In total, there were 130 such articles for the month, of which only 52, fewer than half, were included in the topic "NFL/Kaepernick" used to generate the time series plot in Figure 5.25. However, 99 of these 130 articles belong to the topic labeled "Religion/Divisive Issues", with 30 articles belonging to both topics. The time series for the "Religion/Divisive Issues" topic is presented in Figure 5.26. The labels for the topics that match the "Religion/-Divisive Issues" topic from September 2016 contain terms such as "Religion/Discrimination", "Religion/Books," "Racism/Discrimination," "Religion/Social Issues," and "Families/Lifestyles" to list the most common occurrences.

The value for September 2016 is consistent with the other months included in this time series. The initial reaction to Colin Kaepernick kneeling during the national anthem was discussed in such a way as to differentiate most articles on this subject

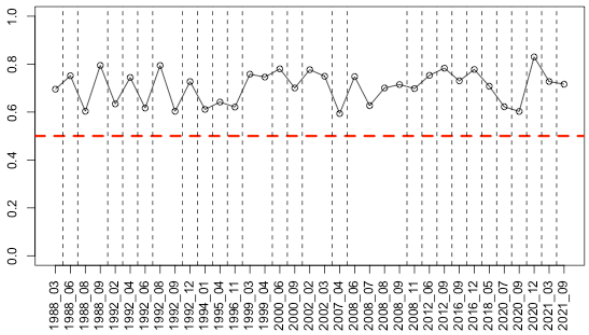


Figure 5.26: Time series for the "Religion/Divisive Issues topic from September of 2016.

from other sports-related articles. The articles that are not sports-centered and discuss Kaepernick fit into an existing topic, which generally has a slightly less positive sentiment than the NFL/Sports topics presented in Figure 5.25. Thus, the polarization generated from this event did not affect the global NFL topic as anticipated. Although the event involved an NFL player, the discussion was closer to an already slightly more polarized topic.

5.3.3 International News

Few topics have erratic shifts in polarization that appear in the time series for articles related to international news. As evidence to support this claim, we present the time series for the topic labeled "Afghan/Iraq Wars/International" from June of 2012 in Figure 5.27. The majority of topics that matched the "Afghan/Iraq Wars/International" topic from June of 2012 were labeled "International", with a few having

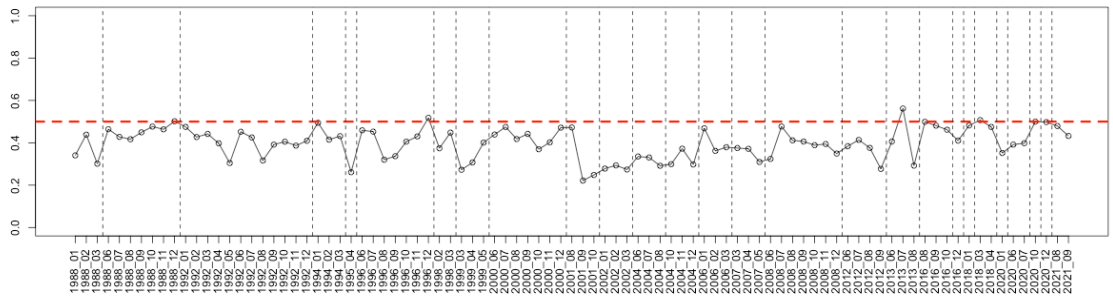


Figure 5.27: Time series plot for the Afghan/Iraq Wars/International from June 2012.

added descriptors such as "War on Terror," "Snowden" (for Eric Snowden), and "Egyptian Uprising". Four topics were labeled "US Politics", a possible misnomer by the author. An outlier, in terms of labels, is the topic from May of 1992, labeled "LA Riots/Bosnian War/Black May in Thailand". The LA Riots refer to the riots that occurred from April 29th to May 4, 1992, following the acquittal of the four police officers charged with excessive force in the arrest and beating of Rodney King. "Black May in Thailand" refers to a series of mass protests and subsequent crackdowns by security forces and police in Bangkok, Thailand, during May 1992.

The volatility inherent in such a topic about international events is not unexpected. Comparatively speaking, the number of articles discussing international events is significantly lower than the number of articles discussing domestic events. Thus, not all international newsworthy events receive coverage to the same extent as expected if a similar event occurs within the United States. Additionally, international news changes rapidly, and discussions surrounding such news can change just as rapidly,

as evidenced by the sudden changes in percent positive from adjacent months.

Figure 5.27 contains several interesting changes in percent positive. One notable change is the shift observed from August 2001 to September 2001. The topic from August 2001 was labeled "US Politics" but contains keywords such as "Palestine," "NATO," "Taliban," "Israel," and "Arafat." Thus, this topic is related to international news and appears mislabeled by the author. In either case, the topic from August 2001 is highly polarized with a percent positive value of 0.47. Compared with the following month, September 2001, we see a significant drop in polarization since the percentage positive value for this month is 0.22. Given the events of September 11, 2001, the sudden drop in positive sentiment for an international topic is not surprising. At first glance, it might be surprising that the discussion around this topic became less polarized, given the tragic events that occurred that month. However, the events of 9-11 had a bridging effect, bringing together disparate groups united by a shared tragedy. A similar phenomenon occurs with a topic from April 1995. On April 19, that year, Timothy McVeigh and Terry Nichols detonated a bomb in front of the Alfred P. Murrah Federal Building in Oklahoma City, Oklahoma. Although much like 9-11, the events of that day were tragic and horrendous, nevertheless, they had a bridging effect. The discourse around such an event is inherently negative, and we see that in both instances, which explains why most of the sentiment around such topics is negative. However, there is still significant agreement on this negative sentiment, indicating a low-polarized environment/topic.

On the contrary, an event that lacked the bridging capital that occurred following 9-11 and the OKC bombing was the leak of documents by Edward Snowden in late

June 2013. Almost instantaneously, news of the leaks and the surveillance program detailed in the leak sparked a public debate on the trade-off between security and privacy. Snowden became a contentious talking point, with some people ready to condemn him and others ready to praise him. This division is exactly what we see in the time series plot for July 2013, the first full month following the leak. The topic that month was labeled "Egyptian Uprising/Snowden/International" and has the highest positive percentage value of any observation in this time series and is one of only four observations in the series that obtains a percent positive greater than 0.5. These events greatly impacted the discourse around a topic with an inherently negative sentiment, and we can see clear evidence of the polarization that follows such events.

The three examples presented in this section are the tip of the iceberg regarding possible events and topics one could explore in this framework and the available data. However, a deep dive into each of these topics is beyond the scope of this document. We aim to demonstrate the effectiveness of our method in detecting periods of high and low polarization. Thus, we divert our attention from single topics and the detection of polarization towards investigating potential differences between newspapers, their coverage of events, and the discourse used to discuss said events.

5.4 Newspaper Comparisons

As mentioned in Section 4.1, we included 20 publication sources in the NewsBank data set. An early goal for this analysis was to determine the differences between these publications. For example, are there regional differences in the discussion around various topics? Additionally, many papers are owned by the same ownership groups, which naturally raises several questions, such as "Are topics discussed differently by different ownership groups?" and "Do publications owned by the same ownership group discuss topics similarly, or are they distinct?" In this section, we investigate the differences between newspapers and answer as many of these questions as possible.

5.4.1 Regional Differences in Global Sentiment

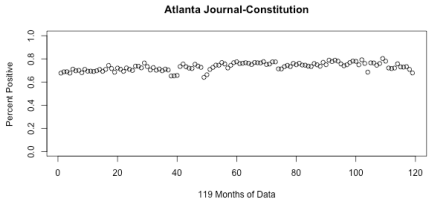
Given that the United States has several diverse cultural regions spread across all fifty states, a logical extension of our current investigation of polarization is to inquire about any regional differences that could exist between the included newspapers. Figure 4.2 shows the geographic distribution of the 18 regional papers selected for analysis. Two of the papers included, USA Today and Christian Science Monitor, are national papers. The number of ways to partition these papers into reasonable regions is too numerous to list. Thus, instead of attempting to justify a particular partition, we opted to compare the time series data for each publication and explore if similarities between publications would "map" onto a reasonable regional partition.

To begin with, we compared all articles based on all months and all articles available each month. Figure 5.28 presents the overall percent positive for each paper. There are scores of interesting observations about these plots, such as the clear change point that occurred during the mid-2010's at the Chicago Sun-Times. However, investigating these observations is beyond the scope of this dissertation. However, these time series give us our first basis for comparing each paper.

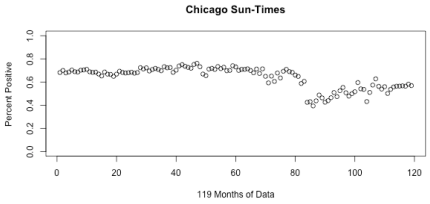
Each publication has a slightly positive sentiment when considering all articles. As 0.5 is our target percent positive for polarization, using this as our base to compare publications has intuitional appeal. For comparison, we computed the mean absolute deviation of each observation in each respective time series. The exact formal for this "deviation metric" is given below in Equation 5.1, where $x_{p,m}$ represents the overall percent positive for publication p during month m and n_p is the number of months we have data for publication p.

$$deviation\_metric = \frac{1}{n_p} \sum_{m=1}^{n_p} |x_{p,m} - 0.5| \tag{5.1}$$

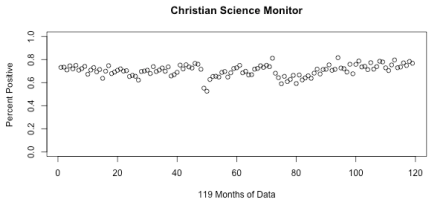
In order to cluster the newspapers, we compute the $deviation\_metric$ for each paper and observe how the papers cluster according to this metric. The distribution of this metric for each of the 20 publications is shown in Figure 5.29. This distribution has three modes, which suggests three clusters. The first cluster consists of papers with a $difference\_metric$ less than 0.16, the second cluster is papers with values between 0.16 and 0.2, and the final cluster is papers with values greater than 0.2. Cluster one contains the papers Syracuse Herald-Journal, Miami Herald, Seattle



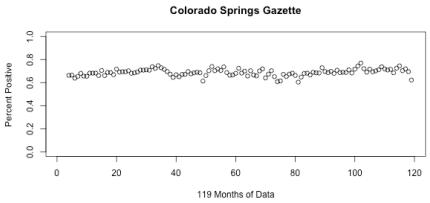
(a) Atlanta Journal-Constitution



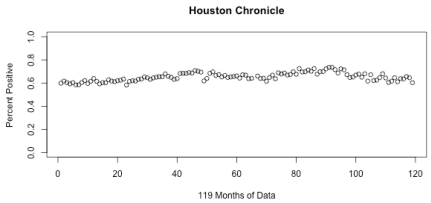
(b) Chicago Sun-Times



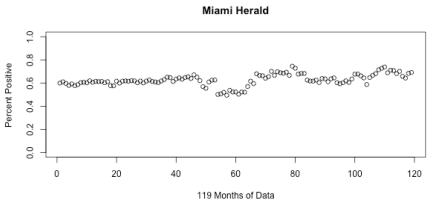
(c) Christian Science Monitor



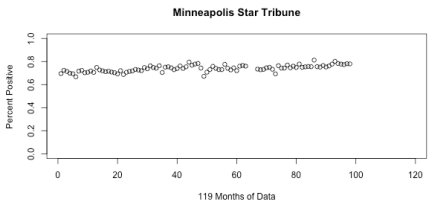
(d) Chicago Sun-Times



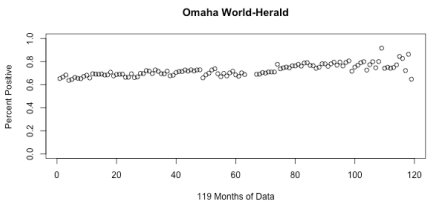
(e) Houston Chronicle



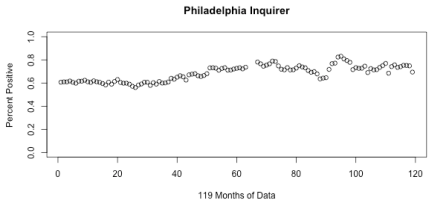
(f) Miami Herald



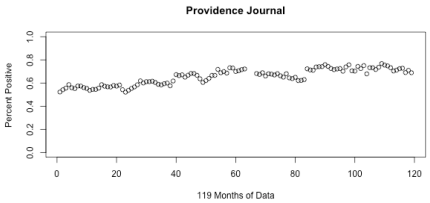
(g) Minneapolis Star Tribune



(h) Omaha World-Herald

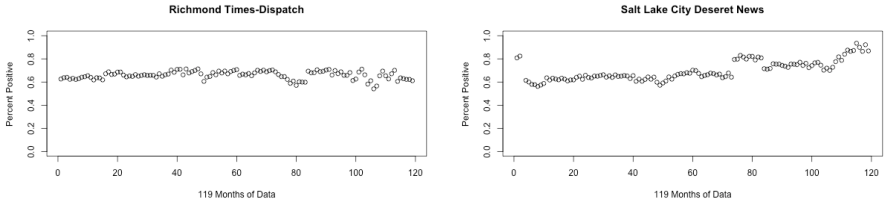


(i) Philadelphia Inquirer



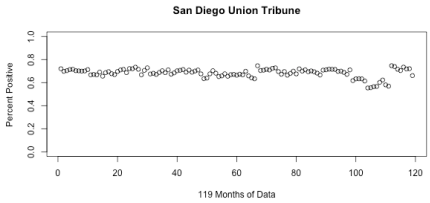
(j) Providence Journal

Figure 5.28: Percent of articles with positive sentiment from each publication.

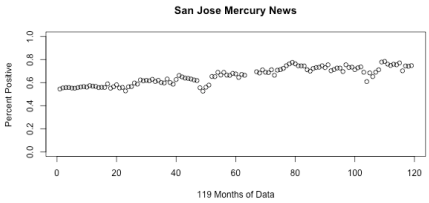


(k) Richmond Time-Dispatch

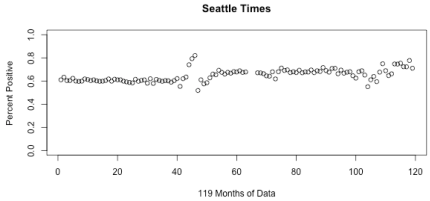
(l) Salt Lake City Deseret



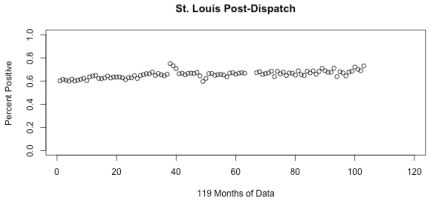
(m) San Diego Union Tribune



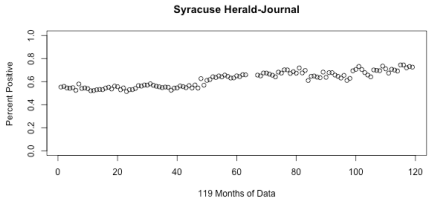
(n) San Jose Merury News



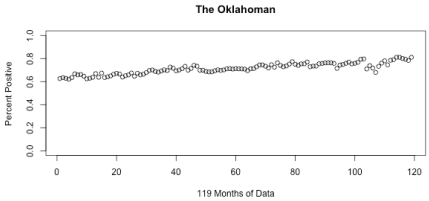
(o) Seattle Times



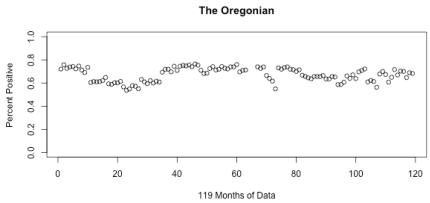
(p) St. Louis Post-Dispatch



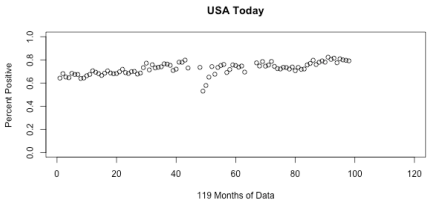
(q) Syracuse Herald-Journal



(r) The Oklahoman



(s) The Oregonian



(t) USA Today

Figure 5.28: Percent of articles with positive sentiment from each publication (cont.).

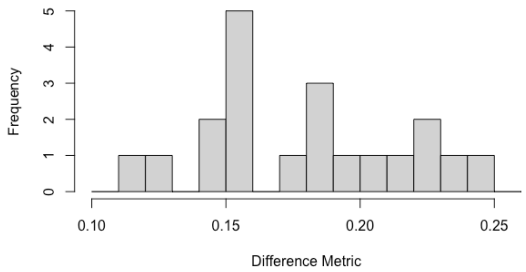


Figure 5.29: Distribution of $difference\_metric$ for the 20 publications from NewsBank.

Times, Chicago Sun-Times, Houston Chronicle, Providence Journal, St. Louis Post-Dispatch, San Jose Mercury News, and the Richmond Times-Dispatch. The second cluster contains The Oregonian, San Diego Union-Tribune, Philadelphia Inquirer, Colorado Springs Gazette, and the Salt Lake City Deseret News. The third cluster comprises Christian Science Monitor, The Oklahoman, Omaha World-Herald, USA Today, Atlanta Journal-Constitution, and the Minneapolis Star Tribune. Other than the two national papers ending up in the same cluster, there does not appear to be any regional association among the clusters. The lack of regional association is not surprising, but our real goal is to determine any regional differences based on the context of specific topics.

5.4.2 Topic Specific Regional Differences

Given the political, cultural, and ethnic diversity in the United States, it is reasonable to anticipate these differences in the discourse of various topics, especially those

that touch upon these differences. Additionally, many of these diversity traits are correlated with the location/region of the United States, which gives a reasonable expectation that discourse of various topics changes across regions. In order to investigate this possibility, we selected several topics of interest and compared the distribution of a difference metric similar to the one given in Equation 5.1. However, we remove the absolute value for the topic-specific differences and compare the mean deviation from 0.5 of each entry in the time series. The exact formulation is given in Equation 5.2.

$$topic\_deviation\_metric = \frac{1}{n_p} \sum_{m=1}^{n_p} (x_{p,m} - 0.5) \tag{5.2}$$

As an initial investigation, we consider the "Iraq War/International" topic from October 2004. This is more than three years after the attacks on 9-11, and the bridging that developed between "groups" following the events of that day has subsided. The Iraq war has been ongoing for 19 months, and the US is one month away from an election that will decide the fate of this war. In total, 68 topics match the topic under consideration, with labels such as "International", "Iraq/Afghan Wars," "War on Terrorism," and "Middle East." The results of applying the metric in Equation 5.2 in each of the twenty newspapers are shown in Figure 5.30 and listed in Table 5.3.

There appears to be a clear separation between the Richmond Times-Dispatch, with a value of -0.086, and the San Diego Union-Tribune, with a value of -0.04, splitting the papers into two clusters. The left cluster consists of the ten regional papers:

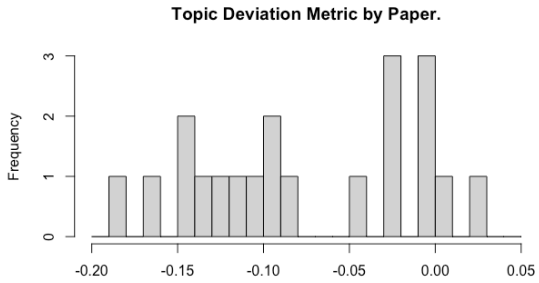


Figure 5.30: Topic deviation metric for each paper centered around 0.5 for the topic "Iraq War/International" from October of 2004.

St. Louis Post-Dispatch	-0.18
Seattle Times	-0.167
Houston Chronicle	-0.145
Philadelphia	-0.143
Miami Herald	-0.137
San Jose Mercury News	-0.128
Salt Lake City Deseret News	-0.112
Chicago Sun-Times	-0.108
Atlanta Journal-Constitution	-0.094
USA Today	-0.091
Richmond Times-Dispatch	-0.086
San Diego Union Tribune	-0.04
Syracuse Herald-Journal	-0.028
Christian Science Monitor	-0.021
Providence Journal	-0.02
Minneapolis Star Tribune	-0.009
The Oregonian	-0.005
Colorado Springs Gazette	-0.0004
The Oklahoman	0.009
Omaha World-Herald	0.024

Table 5.3: Deviation metric for each of the twenty papers.

St. Louis Post-Dispatch, Seattle Times, Houston Chronicle, Philadelphia Inquirer, Miami Herald, San Jose Mercury News, Salt Lake City Deseret News, Chicago Sun-Times, Atlanta Journal-Constitution, and Richmond Times-Dispatch, as well as the national paper USA Today. The right cluster contains the eight regional papers San Diego Union-Tribune, Syracuse Herald-Journal, Providence Journal, Minneapolis Star Tribune, The Oregonian, Colorado Springs Gazette, The Oklahoman, and Omaha World-Herald, along with the national paper Christian Science Monitor. Any regional association among these clusters is tenuous at best, and the clusters appear to separate many of the papers that one would anticipate appearing in the same regions. For example, the Pacific Northwest papers Seattle Times and the Oregonian are in different clusters, the West Coast papers San Jose Mercury News and San Diego Union-Tribune, and the Rocky Mountain papers Salt Lake City Deseret News and the Colorado Springs Gazette. These are just a few of the "pairings" one would expect in any regional partition of the given papers that are not present in the current clusters.

Another time and topic that might have a regional effect on polarization is the 2000 US presidential election. This election was between then Governor of Texas George W. Bush and then Vice President Al Gore. It was one of the closest presidential elections in recent US history, with an electoral college vote count of 271 for Bush and 266 for Gore, with the state of Florida ultimately deciding the contest. Bush's margin of victory in Florida was certified to be 537 votes. This was the first election since 1888 in which the winner of the electoral college vote lost the national popular vote. This was a very close election, with support for each candidate localized to

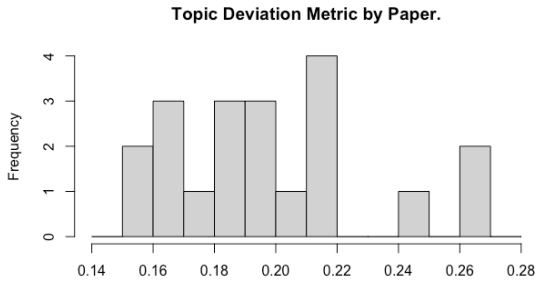


Figure 5.31: Topic deviation metric for each paper centered around 0.5 for the topic "US Politics" from November of 2000.

specific regions. For example, the states Gore won are located in the Northeast, Northern Midwest, and the West Coast, with Bush winning all the Southern states, Southern Midwest, the Plain states, and the Rocky Mountain states. Given such a regional effect in terms of which states supported a given candidate, it is possible that such an effect impacted the sentiment regional papers used when discussing this topic. A histogram of the deviation metric for this topic is presented in Figure 5.31. In total, 57 months contained topics that matched the chosen "US Politics" topic from November of 2000. Most of these matched topics have been labeled "US Politics," with a few labeled "Election/Voting" or some combination of these words.

As we can see from Figure 5.31, most of the papers contain deviation metric values between 0.15 and 0.22. There are three potential outliers with values of 0.24, 0.262, and 0.269. These are the Omaha World-Herald, the Minneapolis Star Tribune, and the Christian Science Monitor. Once again, we have very little, if any, regional effect on the discourse surrounding the 2000 US Presidential election topic.

Given the sheer volume of months that contained a topic that matched the chosen

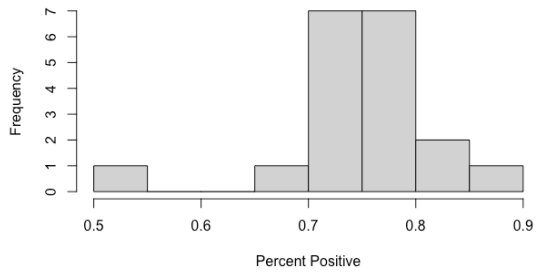


Figure 5.32: Percent Positive for each of the 20 newspapers for the topic "US Politics" from November of 2000.

"US Politics" topic from November 2000 (57 out of a possible 119 months), any regional effect related to the outcome of this election is lost in the aggravation of all 57 months, which run from January 1988 to April 2021. These 57 topics span over three decades and while they are all related in terms of words, there is still a diversity of events and sentiments associated with these events. Thus, we examine the single month of November 2000 and compare the percent positive of each newspaper. These values are presented in Figure 5.32.

The distribution of percent positive values is unimodal and reasonably compact. There is possibly an outlier, The Seattle Times, with a percent positive value of exactly 0.5. The reason for such a well-rounded number is that only four articles from this publication are in the given topic. Two of these articles have an overall positive sentiment, and two have an overall negative sentiment, and drawing any significant conclusions from such a small sample size is fraught with peril. Clearly, there is no regional effect on the percent positive for the given topic.

As a final example of our investigation into the regional effects on specific topics,

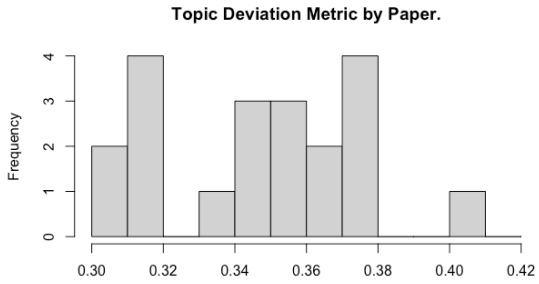


Figure 5.33: Topic deviation metric for each paper centered around 0.5 for the topic "NBA" from June of 2016.

we present the "NBA" topic from June 2016. As we have seen, sports topics are typically unpolarized and have a very positive sentiment. We selected this specific topic because of the events that occurred during the NBA Finals that year. It was a rematch between the defending champions, the Golden State Warriors, and the Cleveland Cavaliers. Going into the series, the Warriors were favored, having set the regular season win record with 73 and overcome a 3-1 deficit in the best of 7-game Western Conference Finals against the Oklahoma City Thunder. This series was one of the more contentious NBA Finals in recent memory due in no small part to the Cavaliers coming back to win the 7-game series after being down 3-1, the first team in NBA history to do so in the NBA Finals. Fifty-two other monthly topics matched the given topic, all sports-related. Figure 5.33 shows the distribution of deviation metric values.

Again, we see that most papers obtain a deviation metric value in a compact interval, in this case between 0.3 and 0.38. There is a possible outlier, the Christian Science Monitor, with a value of 0.409. This is unsurprising as the Christian Science

Monitor rarely reports on sports stories. For example, during the entire month of June 2016, the Christian Science Monitor published one article labeled "NBA". Yet again, we find no regional effect on this topic as there is general agreement among the newspapers on the tone used to discuss the said topic.

We handpicked the above three topics to investigate potential examples of regional effects on specific topics and in each case we were unable to detect any difference. In addition, we investigated several more topics and no apparent regional effect was present for each topic. Given these results, we move our analysis away from regional effects and focus on known differences between papers, specifically political endorsements for president and differences between papers owned by the same entity.

5.5 Comparing Polarization of Papers Based on Endorsements and Ownership

In the previous section, we attempted to discover regional differences in polarization by looking for clusters in various polarization time series for each paper. In this section, we again look for polarization differences between papers; however, in this case we will use known associations between papers. This inquiry will center around two known quantities of each paper, which presidential candidate the paper endorsed for each of the presidential elections available, and the ownership groups of each paper.

5.5.1 Endorsement Based Analysis

Newspapers have been supporting political candidates for well over 150 years. In 1860, the New York Times supported "Mr. Lincoln, of Illinois, familiarly known as 'Old Abe', age 51, height six feet seven, by profession Rail-Splitter." Although political endorsements by a newspaper do not fully represent the paper's political views and leanings, endorsements offer a clear partition that is, at least in part, driven by the political ideology of the higher-ups within the paper. We collect the specific endorsements for each article from the following websites: [Wik], [Bar], and [App].

In order to evaluate any differences between papers based on endorsements, we utilize the same polarization time series plots. However, we only consider topics that appear around the time of a given election to make a comparison. As an illustrative example, we present a topic that appeared during one of the closest Presidential Elections in recent memory, the "Economy" topic from December of 2000. During this election cycle, the newspapers Providence Journal, Syracuse Herald-Journal, Seattle Times, The Oregonian, San Diego Union-Tribune, Richmond Times-Dispatch, Houston Chronicle, The Oklahoman, Omaha World-Herald, and the Chicago Sun-Times all endorsed George W. Bush. The San Jose Mercury News, Philadelphia Inquirer, Miami Herald, St. Louis Post-Dispatch, and Minneapolis Star-Tribune papers endorsed Al Gore. The remaining papers did not endorse a candidate except for the Atlanta Journal-Constitution, which were two separate publications at that time, the Atlanta Journal and the Atlanta Constitution. The Atlanta Journal endorsed

Bush, and the Atlanta Constitution endorsed Al Gore. Unfortunately, NewsBank does not distinguish between the two publications for historical articles, and thus we exclude articles from both publications from our analysis. The time series for all the monthly topics that match with the topic labeled "Economy" from December 2000 is shown in Figure 5.34.

While our primary focus is on the period surrounding the 2000 election, we include all monthly topics that matched to provide a baseline and potentially illustrate any divergence that occurs during the appropriate period. This divergence is exactly what we observe in Figure 5.34. All three time series are in lockstep in the available months before October of 2000. There are slight deviations between each series, but these differences pale in comparison to the differences we saw from October 2000 through February 2002. In addition, we see a convergence of the series until the early 2010s, following the divergence in the early 2000s. It is interesting to note that the period between October 2000 and February 2002 contains the terrorist attacks of September 11. These attacks occurred while George W. Bush was president, and we see a clear difference in the sentiment used by newspapers that supported then-candidate Bush during the 2000 election and the papers that either supported candidate Gore or did not endorse any candidate.

Another example of a divergent time series based on candidate endorsement comes from the "International" topic from November 1988. This election was between then Vice President George H. W. Bush and Michael Dukakis, the Governor of Massachusetts. During this election, the Syracuse Herald-Journal, The Oregonian, San Diego Union-Tribune, Richmond Times-Dispatch, Houston Chronicle, The Ok-

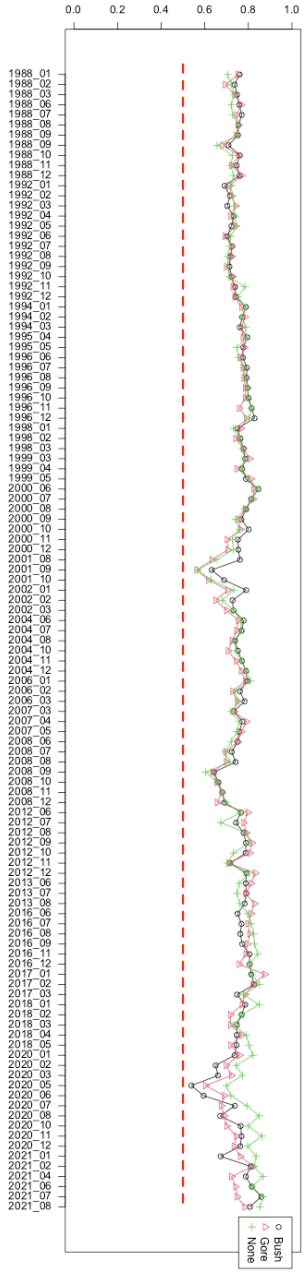


Figure 5.34: Polarization time series for the papers which endorsed Bush, Gore, and no Candidate.

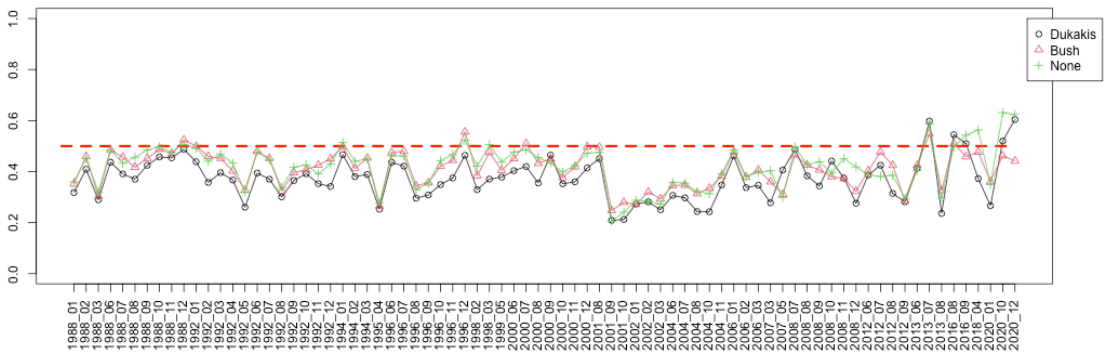


Figure 5.35: Three polarization time series for papers that supported George H. W. Bush, Michael Dukakis, and did not support either candidate in the 1988 Presidential Election.

lahoman, Omaha World-Herald, Salt Lake City Deseret News, and Chicago Sun-Times all supported George H. W. Bush. The Seattle Times, Philadelphia Inquirer, Atlanta Journal-Constitution, St. Louis Post-Dispatch, and Minneapolis Star Tribune all supported Michael Dukakis. The remaining papers from our 20 selected did not endorse. The time series for all topics that match the "International" topic from November 1988 is presented in Figure 5.35.

Figure 5.35 tells a different story from the one presented in Figure 5.34. Here, we have distinct time series right away, and they stay separate for the majority of the sequence. What is interesting about this plot is that the series for papers that endorse Dukakis is consistently lower than the series for the papers that endorse Bush. This implies that the papers that supported Dukakis routinely used a more negative sentiment when compared to papers that endorsed Bush when discussing international news. Of course, we must be careful when drawing conclusions based on this observa-

tion. As stressed in every STAT 101 class ever, "Correlation is not Causation," and it could simply be the case that the worldview that led the higher-ups at papers to endorse Dukakis also caused them to be more pessimistic about international news or vice versa for papers that endorsed Bush. The critical observation is that there is a difference between these sets of papers and how they discuss this topic.

Of course, the plots presented in Figures 5.34 and 5.35 show a difference; most of the topics investigated along the endorsement partition do not differ. In total, we explored 46 topics that appeared around each election and less than a quarter showed any difference in the time series near the time the endorsements were published. For example, consider the "US Politics" topic from November 2012. This election involved then-President Obama and Mitt Romney. Seven papers endorsed Obama, and five endorsed Romney, leaving eight papers that did not make an endorsement. Figure 5.36 shows the time series for the matched topics. These series are consistently correlated and diverge only years after the 2012 election. Suggesting that the cause of these differences is other factors in addition to which candidate the article endorsed in 2012.

Although there are differences between topic polarization based on political endorsements, these differences are not consistent from one election to the next or between topics related to politics. Political endorsements may not be a strong enough signal to indicate polarization with regard to newspaper articles.

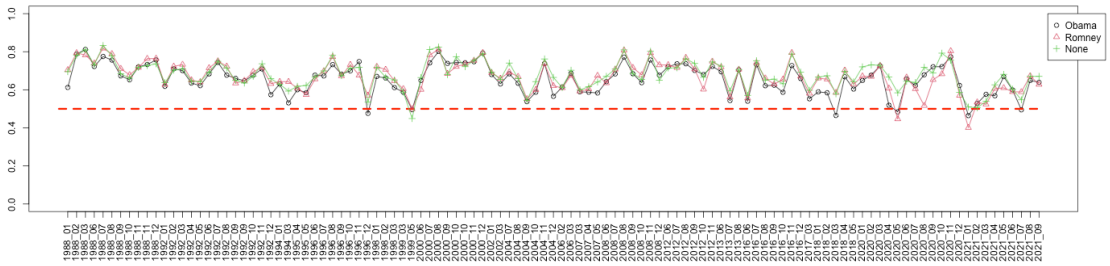


Figure 5.36: Three polarization time series for papers that supported Barack Obama, Mitt Romney, and did not support either candidate in the 2012 Presidential Election.

5.5.2 Ownership Based Analysis

The final comparison among newspapers is to partition based on ownership group. Among our 20 selected publications, four ownership groups own multiple papers. Advance Publications, Inc. owns the Syracuse Herald-Journal and The Oregonian. Alden Global Capital owns the San Diego Union-Tribune and the San Jose Mercury News. Gannett Co., Inc. owns the Providence Journal, USA Today, and The Oklahoman. Finally, Lee Enterprises, Inc. owns the Richmond Times-Dispatch, the Omaha World-Herald, and the St. Louis Post-Dispatch. The owners of each of the remaining papers are groups that own only one paper in our data. For this analysis, we construct a time series by grouping the above papers based on ownership group; the remaining papers are grouped and included in the time series plots and "Rest". This gives a background reference by which to compare the other series. Because many of these papers were purchased relatively recently, we narrowed our focus to only include the most recent topics, with the twenty topics investigated only going back as far as March of 2018. We still display the entire time series for context.

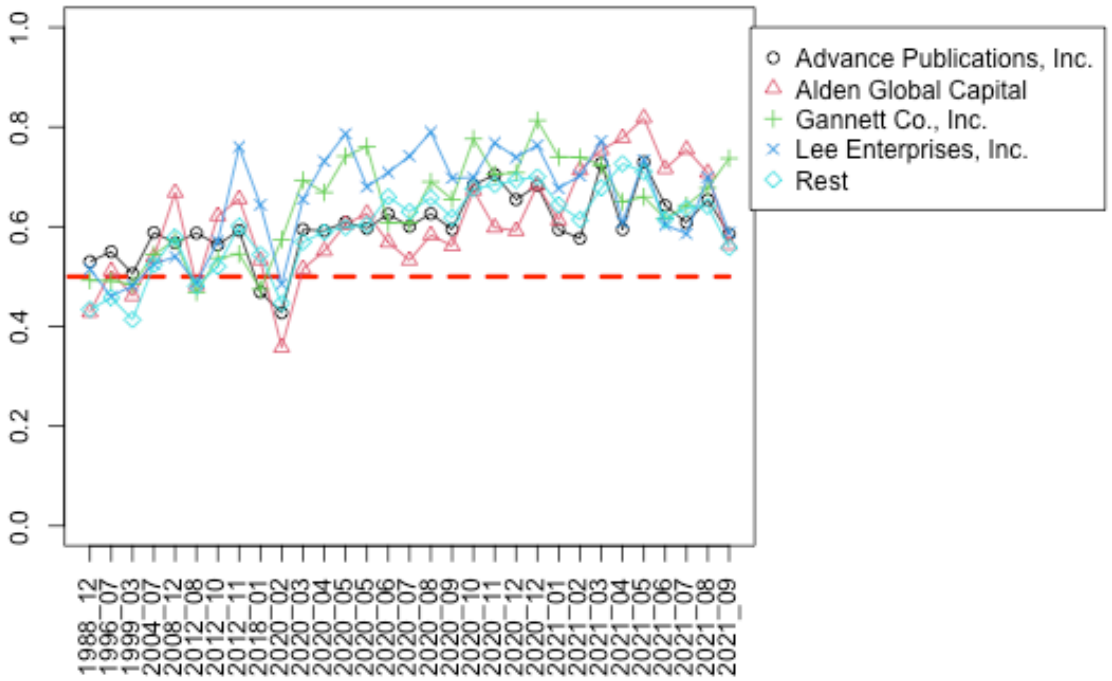


Figure 5.37: The time series for the four ownership groups that own multiple papers in our data set and the time series for all the remaining papers, which are aggregated together, are based on selecting topics that match the "COVID" topic from September 2021.

Our initial illustrative topic is labeled "COVID" and comes from September 2021, the latest month we have data. The time series for this topic and all the ones that matched are shown in Figure 5.37. Most of these topics are from February 2020 or later. However, there are some topics that match from pre-2020, all of which were labeled "Healthcare," which is consistent with a COVID topic.

During the initial months of the COVID lockdown, March 2020 through December

2020, there is a wide discrepancy between the newspapers with the highest percentage of positive sentiment and those with the lowest percentage of positive sentiment. During this period, Lee Enterprises, Inc. group has the highest positive sentiment, followed by Gannett Co., Inc., Advance Publications, Inc., and finally, Alden Global Capital has the lowest percent positive. During this period, Lee Enterprises' papers have an average percent positive of 0.72, while Alden Global papers have an average percent positive of 0.58. This indicates that the Alden authors are more polarized about the COVID pandemic than those at Lee Enterprises or Gannett Co. Interestingly, this dynamic swaps in early 2021 when Alden Global authors present a significant increase in positive sentiment, and the other publications present a significant decrease in positive sentiment. It appears that the Alden Global group became less polarized as the pandemic shifted from 2020 to 2021, while the other publications became more polarized during this period.

Another topic we have considered before but also shows a significant difference between ownership groups is the "US Politics/January 6th" topic from January 2021. The time series for this topic and all the matching topics is presented in Figure 5.38. The most significant disparity between ownership groups occurs during January 2021 and February 2021. During this period, we see a drop in the positive sentiment of all papers. However, the papers owned by Gannett Co., Inc. remain positive, keeping a percentage of positive sentiment above 0.6. The remaining papers all drop below 0.5 percent positive sentiment during one of these two months. This is another excellent illustration of why measures such as entropy or Gini would lose information related to polarization. As mentioned, the ownership group with the largest percent positive

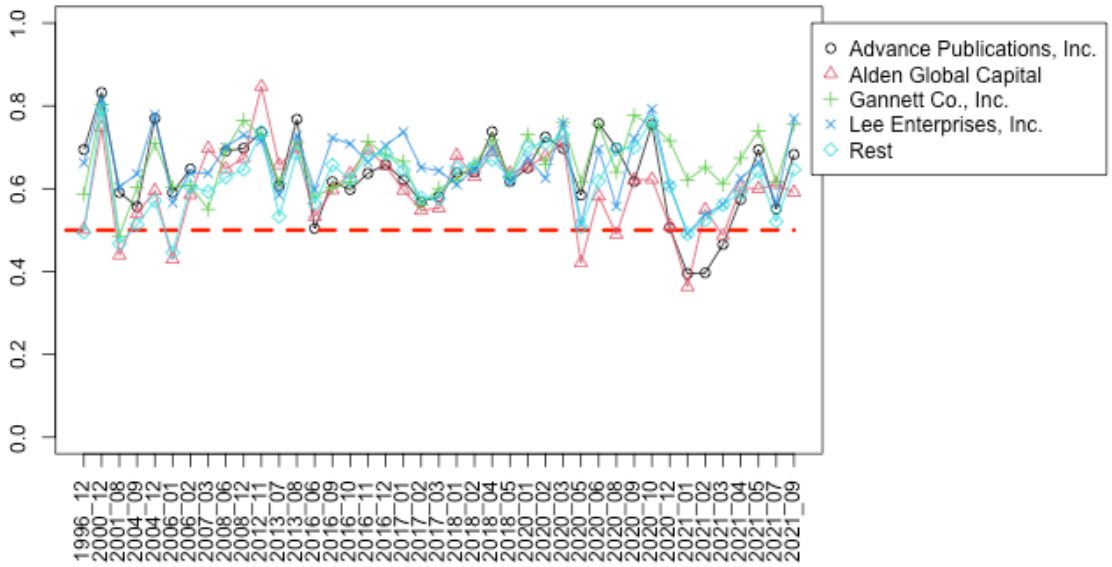


Figure 5.38: Time series for the 4 ownership groups included in our data. The disparity between groups is the highest in the months following the January 6th attack on the Capital.

during January 2021 is Gannett Co., with a value of 0.62. The ownership group with the lowest percent positive during this month is Alden Global 0.36. If we used the entropy measure, these values would correspond to 0.95 and 0.94, respectively. If we used the Gini measure, these values would correspond to 0.47 and 0.46, respectively. Both would indicate that the difference between these papers is negotiable; however, using percent positive, we can see a significant disparity between the groups. Such a disparity indicates a highly polarized topic, as one would expect, given the events of that month.

Evaluating all the possible topics or anticipated polarizing events would be impossi-

ble. A deep dive into these events would require expert knowledge about the events beyond our domain. We present these results to illustrate the effectiveness of our method in detecting polarization.

Chapter 6

Summary and Future Research

In this dissertation, we have introduced a new method to quantify and detect changes in social polarization. This method is a powerful tool that can detect changes in the discourse surrounding events and indicate signatures of polarization. Although we have chosen to focus only on print media in the form of newspaper articles, one could easily apply the method we have developed to other media domains, such as cable news and social media networks. Additionally, this method could be applied to other languages and countries as long as an appropriate sentiment analyzer exists for this context.

While our method focuses on articles as atomic elements for analysis, future research could focus on paragraphs or sentences within articles as atomic elements for analysis. Using articles as atomic elements has the downside of applying the same sentiment value to all of the topics contained within the article. Given that topic models can estimate topic proportions for each article, assigning a topic to paragraphs or even sentences within each article is possible. Dividing articles into their topical components allows for sentiment analysis on each topic represented in an article. This could allow for a more robust detection of polarization on a topic-by-topic

basis, as the sentiment used to discuss different topics within an article could be different.

Finally, a near-limitless number of interesting covariates and time series analysis could be incorporated into our analysis for a given corpus. We gave a glimpse into this possibility by examining differences among newspapers, but this is just the tip of the iceberg. Given enough data, we could consider differences between the authors or sections of the paper. Depending on the context of the corpus, for example, social media posts, we could include covariates that represent engagement of the post and social network interactions. Once these covariates are included, analyzing the resulting time-series by looking for change points could be immensely interesting. The possibilities for discoveries given this quantification method are enormous.

Appendices

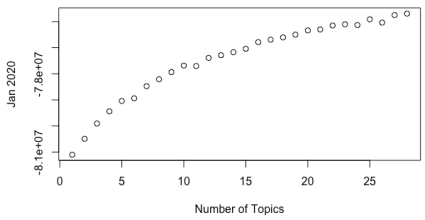
Appendix A

Selecting the Number of Topics

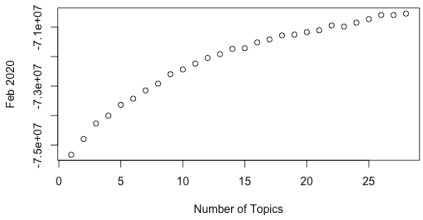
As discussed in Subsection 2.1.1, one thing that topic models have in common is that the number of topics included in the model is a user-defined hyper-parameter, which is not directly estimable from the data set. Cross-validation is a common technique used to justify a particular value for a hyper-parameter. Cross-validation involves selecting an appropriate hyper-parameter space, building models based on the various values of the hyper-parameter in the chosen space, and then comparing each model based on a chosen metric.

The hyper-parameter space for the number of topics is the set of integers greater than or equal to 3 as the stm R package requires a minimum of 3 topics. In order to evaluate the appropriate number of topics for a month's worth of news articles, we took each month of 2020 and fit the stm model to all the articles for each month using $K = 3$ through $K = 30$. Then, using these model fits, we evaluated the log-likelihood function for model fit. The results from these evaluations are plotted in Figures A.1 and A.2.

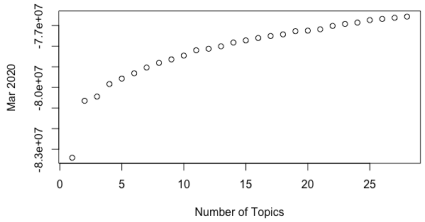
It is well known that the log-likelihood function increases as the number of parameters increases. The traditional method for estimating a hyper-parameter from plots such



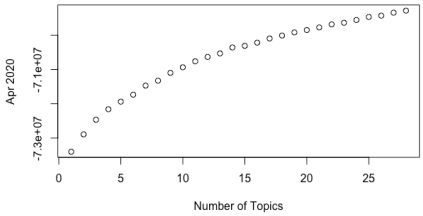
(a) January



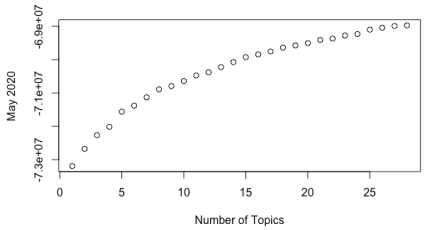
(b) February



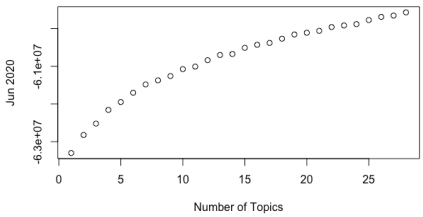
(c) March



(d) April

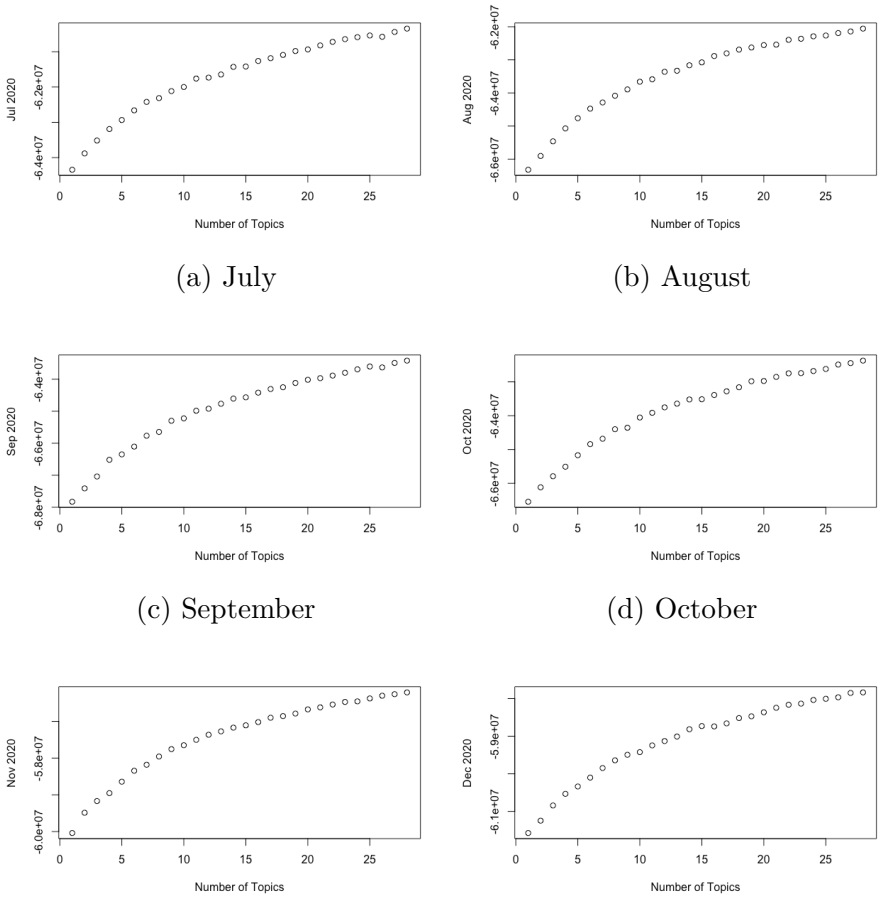


(e) May



(f) June

Figure A.1: Log-Likelihood evaluations for the stm model for each month using $K = 3$ through $K = 30$



(a) July (b) August

(c) September (d) October

(e) November (f) December

Figure A.2: Log-Likelihood evaluations for the stm model for each month using $K = 3$ through $K = 30$

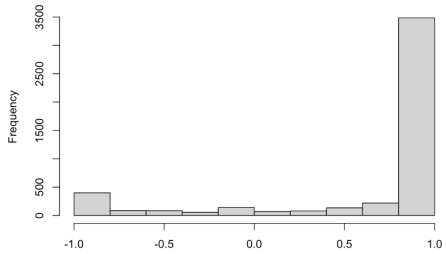
as these is to look for an "elbow" where the amount the log-likelihood function increases changes, and the rate of the increase goes down. While there is not a single K value where this "elbow" occurs for each month, each month appears to have an "elbow" between $K = 8$ and $K = 13$.

Given this information, we made the decision to go with $K = 20$ topics for our model fits. This choice was made to ensure that we capture all of the major topics discussed each month. While it is possible that allowing for topics beyond the 'elbow' point could run the risk of 'splitting' a topic, we deemed this an acceptable risk. We were confident in our topic-matching process and human inspection, which we believed would ensure the accuracy of our method even if we did 'split' a topic. We believed this was a more desirable outcome than the possibility of missing a topic by not allowing enough topics in the model.

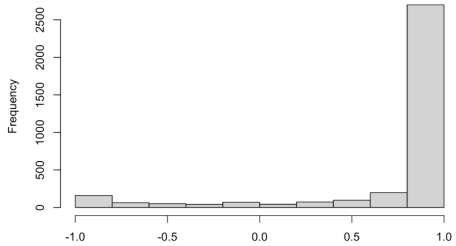
Appendix B

Polarization Plots for the Summer Olympics

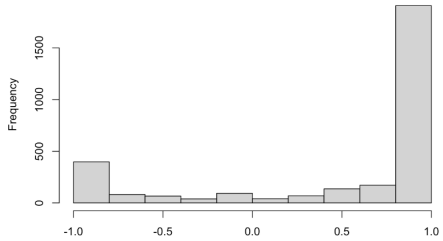
During our initial investigation, we explored the possibility that the Summer Olympics could be a bridging event and lower the overall polarization. We conducted an investigation similar to the one outlined in Section 5.1.2. The resulting polarization plots are remarkably similar to the ones obtained when investigating the Presidential Elections. We present these plots here for the interested reader to compare with the plots from section 5.1.2.



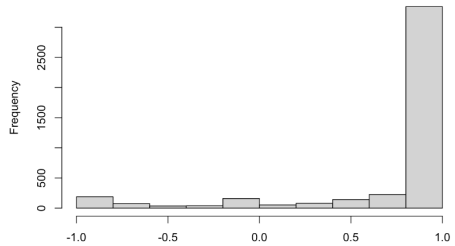
(a) Summer Olympics 1988



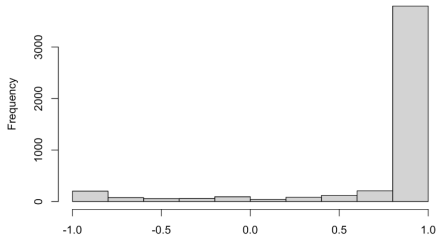
(b) Summer Olympics 1992



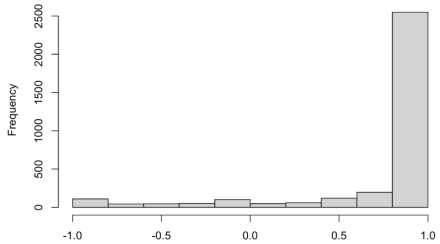
(c) Summer Olympics 1996



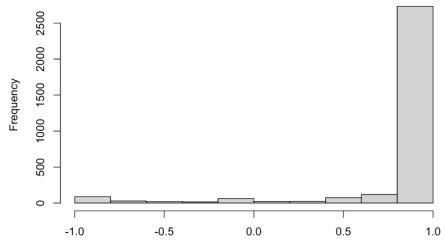
(d) Summer Olympics 2000



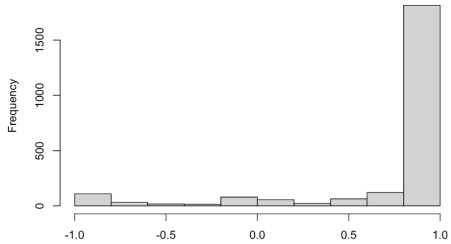
(e) Summer Olympics 2004



(f) Summer Olympics 2008

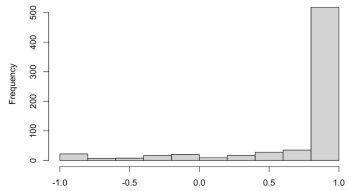


(g) Summer Olympics 2012

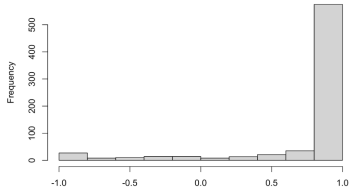


(h) Summer Olympics 2016

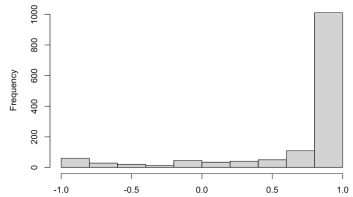
Figure B.1: Polarization plots for all articles within the "Summer Olympics" topic covering the period of each Olympic games.



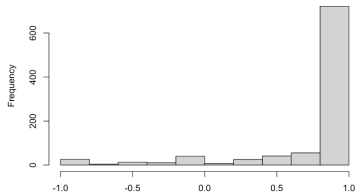
(a) Cooking/Food Topic 1992



(b) Cooking/Food Topic 2000

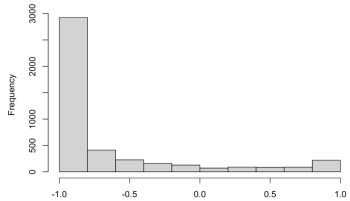


(c) Cooking/Food Topic 2008

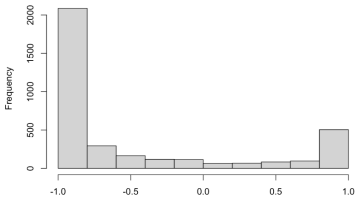


(d) Cooking Food Topic 2016

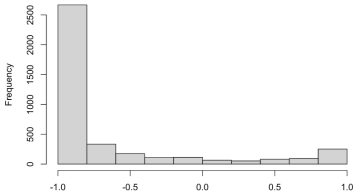
Figure B.2: Polarization plots for all articles within the "Cooking Food" topic covering the Summer Olympics.



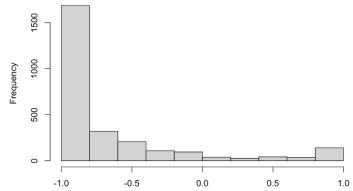
(a) Crime Report Topic 1992



(b) Crime Report Topic 2000



(c) Crime Report Topic 2008



(d) Crime Report Topic 2016

Figure B.3: Polarization plots for all articles within the "Crime Report" topic covering the Summer Olympics.

Bibliography

[Pop35] Tiberiu Popoviciu. "Sur les équations algébriques ayant toutes leurs racines réelles". In: *Mathematica* 9.129-145 (1935), p. 20.

[Dee+90] Scott Deerwester et al. "Indexing by latent semantic analysis". In: *Journal of the American society for information science* 41.6 (1990), pp. 391–407.

[Hof99] Thomas Hofmann. "Probabilistic latent semantic indexing". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999, pp. 50–57.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[BL06a] David Blei and John Lafferty. "Correlated topic models". In: *Advances in neural information processing systems* 18 (2006), p. 147.

[BL06b] David M Blei and John D Lafferty. "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 113–120.

[LM06] Wei Li and Andrew McCallum. "Pachinko allocation: DAG-structured mixture models of topic correlations". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 577–584.

[WM06] Xuerui Wang and Andrew McCallum. "Topics over time: a non-markov continuous-time model of topical trends". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006, pp. 424–433.

[BY09] Humnath Bhandari and Kumi Yasunobu. "What is social capital? A comprehensive review of the concept". In: *Asian Journal of Social Science* 37.3 (2009), pp. 480–510.

[FHT10] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent". In: *Journal of statistical software* 33.1 (2010), p. 1.

[CT11] Chien Chin Chen and You-De Tseng. "Quality evaluation of product reviews using an information quality framework". In: *Decision Support Systems* 50.4 (2011), pp. 755–768.

[Han11] Hahrie Han. *The disappearing center: Engaged citizens, polarization, and American democracy*. 2011.

[KYH12] Hanhoon Kang, Seong Joon Yoo, and Dongil Han. "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews". In: *Expert Systems with Applications* 39.5 (2012), pp. 6000–6010.

[HG14] Clayton Hutto and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of*

the international AAAI conference on web and social media. Vol. 8. 1. 2014, pp. 216–225.

[Qua+15] Xiaojun Quan et al. "Short and sparse text topic modeling via self-aggregation". In: *24th International Joint Conference on Artificial Intelligence, IJCAI 2015*. AAAI Press/International Joint Conferences on Artificial Intelligence. 2015, pp. 2270–2276.

[TAR15] Abinash Tripathy, Ankit Agrawal, and Santanu Kumar Rath. "Classification of sentimental reviews using machine learning techniques". In: *Procedia Computer Science* 57 (2015), pp. 821–829.

[RSA16] Margaret E Roberts, Brandon M Stewart, and Edoardo M Airoldi. "A model of text for experimentation in the social sciences". In: *Journal of the American Statistical Association* 111.515 (2016), pp. 988–1003.

[PMS17] Rajesh Piryani, D Madhavi, and Vivek Kumar Singh. "Analytical mapping of opinion mining and sentiment analysis research during 2000–2015". In: *Information Processing & Management* 53.1 (2017), pp. 122–150.

[Poz+17] Federico Alberto Pozzi et al. "Challenges of sentiment analysis in social networks: an overview". In: *Sentiment analysis in social networks* (2017), pp. 1–11.

[Cam18] James E Campbell. *Polarized: Making sense of a divided America*. Princeton University Press, 2018.

[HL18] Danny Hayes and Jennifer L Lawless. "The decline of local news and its effects: New evidence from longitudinal data". In: *The Journal of Politics* 80.1 (2018), pp. 332–336.

[JJM18] R Johnston, KELVYN Jones, and DAVID Manley. "An increasingly polarized America". In: *Atlas of the 2016 Elections* (2018), pp. 104–110.

[MRS18] Jennifer McCoy, Tahmina Rahman, and Murat Somer. "Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities". In: *American Behavioral Scientist* 62.1 (2018), pp. 16–42.

[RST19] Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. "stm: An R Package for Structural Topic Models". In: *Journal of Statistical Software* 91.2 (2019), pp. 1–40. DOI: 10.18637/jss.v091.i02.

[You+19] Abdallah Yousif et al. "A survey on sentiment analysis of scientific citations". In: *Artificial Intelligence Review* 52 (2019), pp. 1805–1838.

[Haw+20] James Hawdon et al. "Social media use, political polarization, and social capital: is social media tearing the US apart?" In: *International Conference on Human-Computer Interaction*. Springer. 2020, pp. 243–260.

[HL20] Gordon Heltzel and Kristin Laurin. "Polarization in America: Two possible futures". In: *Current opinion in behavioral sciences* 34 (2020), pp. 179–184.

[Kle20] Ezra Klein. *Why we're polarized*. Simon and Schuster, 2020.

[BKB21] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. "A comprehensive survey on sentiment analysis: Approaches, challenges and trends". In: *Knowledge-Based Systems* 226 (2021), p. 107134.

[JPA21] Praphula Kumar Jain, Rajendra Pamula, and Sarfraj Ansari. "A supervised machine learning approach for the credibility assessment of user-generated content". In: *Wireless Personal Communications* 118 (2021), pp. 2469–2485.

[LCT21] Alexander Ligthart, Cagatay Catal, and Bedir Tekinerdogan. "Systematic reviews in sentiment analysis: a tertiary study". In: *Artificial Intelligence Review* (2021), pp. 1–57.

[BGS22] Levi Boxell, Matthew Gentzkow, and Jesse M Shapiro. "Cross-country trends in affective polarization". In: *Review of Economics and Statistics* (2022), pp. 1–60.

[Sat+22] Nicole Satherley et al. "Political attitude change over time following COVID-19 lockdown: Rallying effects and differences between left and right voters". In: *Frontiers in Psychology* 13 (2022), p. 1041957.

[WRK22] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. "A survey on sentiment analysis methods, applications, and challenges". In: *Artificial Intelligence Review* 55.7 (2022), pp. 5731–5780.

[App] Eric M. Appleman. *Endorsements by Newspaper and Magazines*. https://p2000.us/natendorse5.html. Accessed: May 19, 2024.

[Bar] UC Santa Barabra. *The American Presidency Project*. https://www.presidency.ucsb.edu/statistics/data/2020-general-election-editorial-endorsements-major-newspapers. Accessed: May 17, 2024.

[Fac] Science Facts. *Polarization of Light*. https://www.sciencefacts.net/polarization-of-light.html. Accessed: Oct. 26th 2023.

[Lea] Lumen Learning. *Polarization*. https://courses.lumenlearning.com/suny-physics/chapter/27-8-polarization/. Accessed: Dec. 5th 2023.

[Ner] Nerebur. *Dirichlet.pdf*. https://en.wikipedia.org/wiki/File:Dirichlet.pdf. Accessed: Jan. 11th 2024.

[Pro] Committee on Professional Ethics of the American Statistical Association. *Ethical Guidelines for Statistical Practice*. https://www.amstat.org/your-career/ethical-guidelines-for-statistical-practice. Accessed: July 5, 2024.

[Wik] Wikipedia. *Lists of newspaper endorsements in United States presidential elections*. https://en.wikipedia.org/wiki/Lists_of_newspaper_endorsements_in_United_States_presidential_elections. Accessed: May 15, 2024.