




Natural language processing meets spatial time series analysis and geovisualization: identifying and visualizing spatio-topical sentiment trends on Twitter

Hoeyun Kwon ^a, Caglar Koylu ^a and Bryce J. Dietrich ^b

^aDepartment of Geographical and Sustainability Sciences, University of Iowa, Iowa City, IA, USA; ^bDepartment of Political Science, Purdue University, West Lafayette, West Lafayette, IN, USA

ABSTRACT

Previous studies have introduced various approaches for visualizing the spatial and temporal distributions of sentiments expressed on social media. However, many existing methods either overlook the evolving nature of sentiments or fail to account for the spatial distribution of sentiment trends related to specific topics. To gain a comprehensive understanding of how sentiments evolve in relation to topics and geographies, it is essential to capture the dynamic nature of sentiment through time series analysis and geovisualization. This article introduces a workflow that combines natural language processing, spatial time series analysis, and geovisualization techniques to identify and visualize the variations in sentiment trends on Twitter across different geographic regions and topics. By examining the 2016 presidential debates as a case study, we uncover distinct temporal patterns in sentiment distributions across various topics and states. Our findings also show that adjacent states do not always share similar sentiment trends, and that geographic clusters with similar sentiment trends also vary across topics. Failing to consider these variations may result in misunderstanding public discourse and sentiments since they are diverse and dynamic in nature.

KEYWORDS

Sentiment trends; natural language processing; spatial time series analysis; geovisualization; Twitter

1. Introduction

Scholars often attach online communities on social media to physical places, whether it is about elections (Liu et al., 2021), public health (Eichstaedt et al., 2015), simple demographics (Martin et al., 2021), and disasters (Zou et al., 2018). However, we know that social media, by its very nature, often transcends these physical boundaries, and topics and sentiments diffuse among places that are far apart from each other. On social media, millions of individuals express their thoughts and emotions while interacting with one another, which often influences key forms of political behavior, like voting (Key & Heard, 1949; Kinsella et al., 2015). At the very least, this implies that topics and sentiments on social media may be articulated differently in one online community compared to another. Unfortunately, regardless of the level of aggregation (e.g. county, state, or country), previous studies have not fully captured this spatial dynamic, especially when it is combined with temporal trends. Also, despite a few attempts to identify spatiotemporal patterns of topics and sentiments from social media posts (Koylu et al., 2019; Li et al., 2020), existing studies often look at spatial

patterns and temporal trends separately. Separating these two is problematic since the variation in temporal trends is often correlated with spatial patterns. For example, in the context of the 2016 election, certain states were predisposed to vote for Donald Trump based on their past voting behavior. So, when something positive happened to the Trump campaign, these states were more likely to express this sentiment online. Those looking at the aggregated trend may conclude that sentiment was becoming more positive toward Trump during this period, but when decomposed, the time series would show that the shift in sentiment was primarily attributed to only a handful of states. Therefore, to accurately understand the sentiment toward a topic, it is crucial to capture the dynamic nature of sentiment, which further requires the analysis of time series trends of sentiments in relation to subtopics, and geography.

In this study, we introduce a workflow to identify and visualize the spatio-topical sentiment trends on social media by integrating natural language processing with spatial time series analysis and geovisualization. We use the term “spatio-topical sentiment trend” to emphasize the dynamic nature of sentiments that vary across both

geographic and semantic (topical) spaces. Our study makes three contributions. First, our workflow captures the evolution of emerging and disappearing topics along with the associated sentiments, highlighting the temporal variations in sentiments across topical (semantic) and geographic spaces. Second, our workflow helps identify regions that share similar reactions to prevalent topics on social media. We argue that this is indicative of the different types of communities that can form online, but we also think it could help identify relevant blocs of traditional boundaries such as geopolitical regions that could prove helpful for understanding various forms of behavior, like voting. Third, we introduce a novel non-contiguous rectangular cartogram that visualizes spatio-topical sentiment trends (i.e. sentiment trends in relation to subtopics and geographies) and regions that share cohesive reactions to events. Our workflow and cartogram may help scholars identify patterns that would be difficult to unearth without integrating natural language processing, spatial time series analysis, and geovisualization.

To demonstrate our workflow, we use Twitter data related to the United States presidential debates of 2016, which sparked considerable online discourse regarding various political and social issues. These debates have been studied extensively in the literature (Robertson et al., 2019), but little has been written about how the associated online discussions differed in one region or another. Moreover, past work has only used snapshots of online discussions to explore issues related to candidate sentiment (Yaqub et al., 2017). Although we do not look at whether one candidate is preferred over another, our study is the first to identify and visualize spatio-topical sentiment trends, ultimately giving us a better sense of what online discourse looks like regarding these important events. The rise of Donald Trump has received a lot of scholarly attention (Carmines et al., 2016). This study will help shed new light on the subject by looking at how online communities formed around the discussion of Donald Trump and Hillary Clinton shortly after the 2016 presidential debates. While our article centers around the 2016 presidential debates to contextualize our findings, we also expand our scope by conducting a case study on tweets during hurricane Irma in 2017. This additional case study serves as a demonstration of the generalizability and applicability of our workflow to other research domains (see Appendix B).

2. Related work

Retrieving sentiments toward different topics from social media is challenging since one cannot easily

capture and summarize the diverse set of topics and sentiments expressed by many individuals from various geographic areas. Previous work conducted topic and sentiment analysis on social media data to identify public opinions and emotions about events such as natural disasters (Garske et al., 2021; Sit et al., 2019; Yuan et al., 2020), infectious diseases and vaccines (Du et al., 2017; Han et al., 2020; Hu et al., 2021), the stock market (Xu & Keelj, 2014), and elections (Wu et al., 2017; Yao & Wang, 2020). Many studies show that there are spatial and temporal variations in public sentiment and subtopics about an event. For instance, Gruebner et al. (2018) extracted negative emotions from tweets related to Hurricane Sandy and identified variations in their spatial patterns over time. They found that negative emotions were clustered not only spatially but also temporally. Han et al. (2020) extracted topics related to COVID-19, such as seeking medical help, willing to return to work, and praying, and then investigated the spatial and temporal distributions of the social media posts for each topic. Koylu et al. (2019) analyzed temporal and spatial patterns of sentiments toward immigration before and after the Muslim Ban to identify temporal changes in sentiments and subtopics. Although these studies show how social media posts have distinct spatial and temporal patterns of topics and sentiments, they fail to capture sentiment trends that vary across subtopics, geography, and time.

There have been a few attempts to analyze topics and sentiments simultaneously, but they ignore either space or time dimensions. For example, Jeong et al. (2019) analyzed how sentiments varied based on subtopics but did not examine how sentiments varied across space and time. Diakopoulos and Shamma (2010) analyzed the change in sentiment over time in relation to subtopics; however, they did not consider spatial patterns. Li et al. (2020), Koylu (2018), and Koylu (2019) introduced conceptual and analytical approaches to investigate the temporal evolution of topics and their spatial patterns, but neither of the studies considered sentiments toward different topics. Information in social media has multiple dimensions including space, time, topic, and sentiment (Dunkel et al., 2019; Janowicz et al., 2019), and they are all interconnected (Wang & Ye, 2018). Despite this interconnectedness, no attempts have been made thus far to uncover time series trends of sentiments in relation to topics and geography. This is problematic since an accurate understanding of sentiments toward a topic necessitates capturing the dynamic nature of sentiment, which further requires the time series analysis of sentiment trends in relation to topics and geography.

3. Methodology

We introduce an analytical workflow that integrates natural language processing with spatial time series analysis and geovisualization to identify and visualize spatio-topical sentiment trends on Twitter. The term “spatio-topical sentiment trends” refers to the time series trends of sentiments that are embedded in and vary across both geographic spaces and topics. Our goal is to capture the variations in sentiment trends in relation to both sub-topical themes and geographic areas and identify geographic areas that have similar sentiment trends. Our workflow consists of four parts: data processing, natural language processing, spatial time series analysis, and geovisualization (Figure 1). In the following subsections, we explain each component of our analytical workflow.

3.1. Data processing

We collect Twitter data using the Streaming API and a set of keywords related to an event. Among the collected tweets, we only include original tweets written in English by filtering out retweets, duplicates, and non-

English tweets. To capture time series trends of sentiments, our workflow classifies topics and sentiments at the tweet level and computes sentiment trends of tweets by geography and subtopics. However, this process makes the time series data for each geographic unit and each topic become sparser with missing observations at some time points, which makes it difficult or impossible to apply time series analysis. To alleviate the problem of data sparsity, we examine and choose optimal spatial and temporal units of aggregation. In our case study, we use states as the spatial unit of analysis to summarize the temporal trends. Although states are large units of aggregation, some states have fewer Twitter users, which requires combining those states into groups or regions that share similar cultural and socio-political composition.

As public reactions to events on social media are bursty, the number of tweets produced during and immediately after an event corresponds to a large portion of all tweets produced about the event. Bursty reactions to an event usually exhibit an exponentially decreasing distribution, so the number of tweets decreases sharply after each time step. It is another significant challenge in time series analysis: whichever

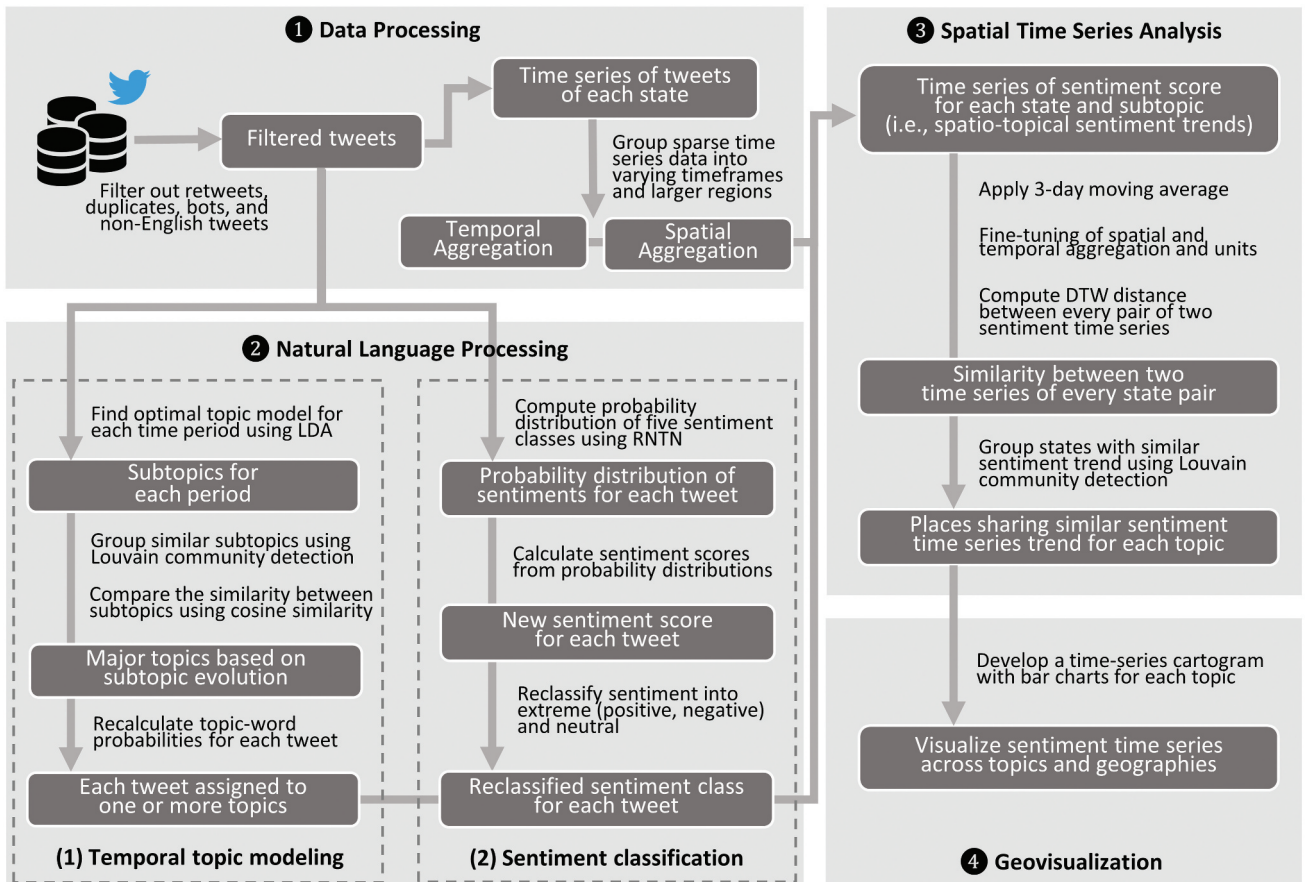


Figure 1. The analytical workflow of this study.

temporal unit is determined to represent the time unit, data further away from the time of the event become sparser. To address this issue, we propose temporal units with varying time lengths that could range from minutes to hours and days, so that time series data are distributed approximately evenly throughout the whole period of analysis. We further discuss the choice of spatial and temporal units and aggregation in [Section 4.1](#), in which we justify our choice with the case study.

3.2. Natural language processing

3.2.1. Temporal topic modeling

We first extract topics for each tweet using Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA estimates a probabilistic distribution of latent topics for each document based on the co-occurrence and frequency of the words in the documents and returns a multinomial distribution that assigns each document a mixture of topics (Blei et al., 2003):

$$P(Z|W, D) = \frac{W_{z+\beta_w}}{\text{total words in } Z + \beta} * D_{Z+\alpha}$$

where $P(Z|W, D)$ refers to the probability that word W from document D falls into topic Z . This probability is calculated by multiplying the normalized frequency of word W in topic Z with the number of words in topic Z in document D . As a result, LDA produces a probabilistic distribution of topics across documents and words across topics. We consider each tweet as a document for performing LDA in this study.

We use a temporal topic modeling approach to capture the temporal evolution of topics. Ignoring temporal evolution is problematic for several reasons. First, the ebb and flow of online discussions mean that a topic could be present in one instance but disappear in the next. Consequently, when the LDA model estimates the topic proportions at the corpus level, that topic would be underrepresented. By incorporating a temporal dimension, our approach does not suffer from the same limitation since topics can be tied to specific combinations of space and time. Second, collective discourse is often diverse and spontaneous, meaning our approach better represents these real-time discussions. Indeed, to think of tweets at the corpus level ignores the way those tweets are actually produced. Instead of thinking of a larger topic distribution encompassing, when people interact online, they are often reacting to an event here and now.

To identify the optimal topic model for each period, we perform a series of topic models for each time period (or an event such as the presidential debates used in this

study). Since topic modeling results vary depending on the number of topics, we experiment with topic models with 5 to 60 topics (i.e. 56 models) for each period that covers an event. For example, in our case study, these events are the three presidential debates. The time period for each event is from the beginning of each debate to 1 week after the debate when the next debate takes place, except the last debate. To determine the optimum number of topics that minimize the number of duplicate topics, we compare the similarity of topics between models of the consecutive number of topics (e.g. such as 5-topic model and 6-topic model). To compare the similarity of topics and identify an optimal model for each time period, we calculate cosine similarity between every pair of two topics from all those 56 models. Cosine similarity measures the cosine of the angle between the two word vectors that form each topic and their word frequencies (Huang, 2008), which ranges between 0 and 1. We follow the rule that the model with a larger number of topics (i.e. 6) should have a cosine similarity value of less than 30% with any topic of the model with a smaller number of topics (i.e. 5). A larger similarity value above this threshold indicates the overlapping of topics between the two models. When there are fewer topics, unique or non-overlapping topics are often combined into fewer topics, whereas a larger number of topics result in topics with overlapping content.

After identifying the optimal topic model for each period, we create a network of pairwise topic similarities to link similar topics within the same time period as well as topics of the optimal models from other periods. [Figure 2](#) illustrates the concept of finding the most similar pairs of topics given a threshold of 0.7 (or 70%) cosine similarity of topics within and between topic models of two different time periods. Using these edges between topics, we construct a network of topic similarity in which a node is a topic, and an edge represents the similarity between a pair of topics. We use the Louvain community detection method to group topics into topic clusters. From now on, we use the term “subtopic” to refer to each topic in a topic model, while we use the term “topic” to refer to “a cluster of topics.”

Grouping subtopics within and across different temporal models is essential because tracing temporal trends requires robust time series of sentiment scores. Without grouping, most emerging topics suffer from sparse time series data, and therefore, it would become impossible to identify their temporal trends of subtopics using time series analysis. Although grouping subtopics into topics (topic clusters) seems to lose the variations within each topic, one can always break each topic into its original subtopics and trace their semantic and temporal patterns. While some topics are formed by a larger

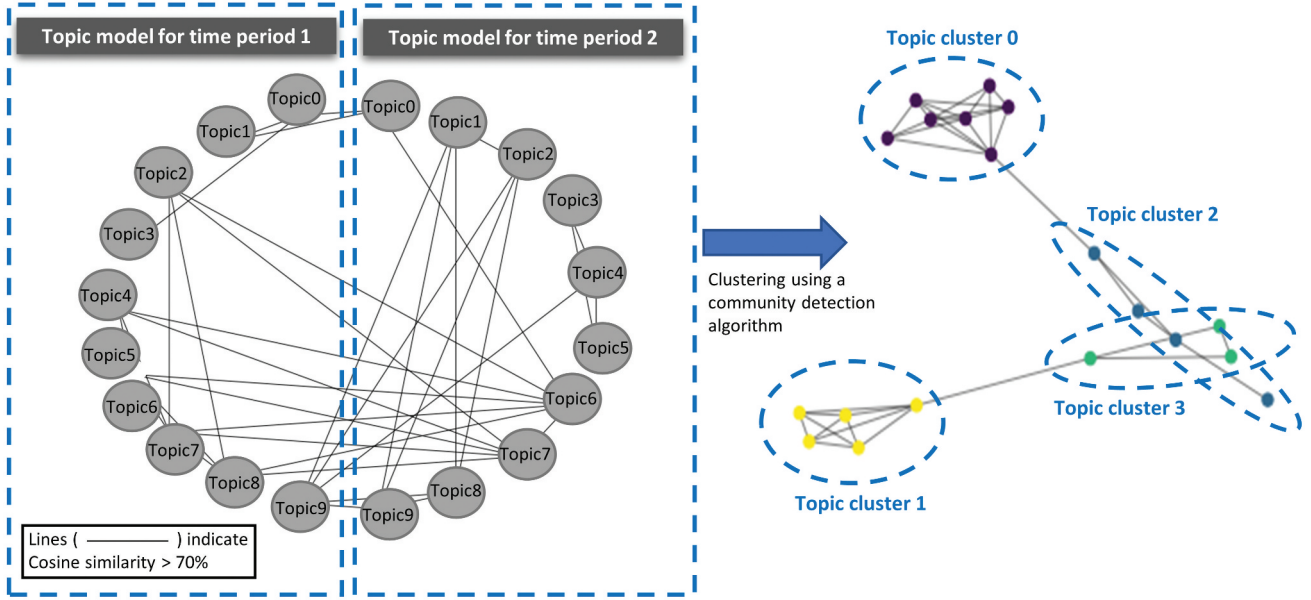


Figure 2. Clustering topics from models of different time periods.

number of subtopics that can be considered topic chains, some other topics may be formed by a single or only a few subtopics because they may represent a short-lived reaction of Twitter users to a particular event. Our topic modeling workflow allows capturing both cases.

We recalculate document-topic and topic-word probability matrices for each tweet to combine the probability of subtopics that fall into the same topic (topic cluster). Finally, we define the topic that each tweet belongs to using a threshold of 15% probability. So, one or more topics can be assigned to each tweet if the cumulative probability of the topic for a given document is above 15%. For example, if the probability of a tweet being included in Topic 1 is 20% and that in topic cluster 3 is 40%, then this tweet is classified into both Topic 1 and Topic 3. The 15% threshold allows us to filter out the noise in the multinomial distribution of topics per document.

3.2.2. Sentiment classification

Sentiment classification allows the identification of people’s emotions, sentiments, and attitudes expressed in natural languages such as tweet texts (Koylu et al., 2019; Parimala et al., 2021). Using sentiment analysis, we can understand the polarity (i.e. positive, or negative) as well as different types of emotions (e.g. happy, angry, and sad) of each text. In this study, we identify the sentiment of each tweet using sentiment analysis using the Recursive Neural Tensor Network (RNTN) model developed by Socher et al. (2013). RNTN uses the Stanford Sentiment Treebank and Recursive Neural

Network, which takes texts of any length as input and returns the probability distribution of five sentiment classes for each tweet: very positive, positive, neutral, negative, and very negative.

Using the RNTN output, we reclassify the sentiment of each tweet into three classes: positive, neutral, and negative, and our approach is significant in two ways. First, we include the neutral sentiment category, unlike many previous studies excluding neutral and focusing on extreme sentiments. It is important to consider neutral because neutral usually makes up a large portion of the sentiment distribution. Thus, excluding neutral may result in removing data excessively and making time series analysis difficult because of data sparsity. Second, we classify the sentiment of each tweet by considering the distinction between positive and negative sentiments instead of using the most prominent sentiment class whose probability is the highest from the RNTN output.

We first compute the sentiment score of each tweet by subtracting the sum of two negative probabilities (negative and very negative) from the sum of two positive probabilities (positive and very positive), simply keeping neutral score as 0. This makes the possible range of new sentiment scores to be between -1 and 1 . For example, there is a tweet A saying “Are you ready to watch the debates?” with the probabilities of [very positive 0, positive 0.1, neutral 0.9, negative 0, very negative 0] and a tweet B saying “That’s a difficult question. I like it!” with the probabilities of [very positive 0.2, positive 0.3, neutral 0.1, negative 0.3, very negative 0.1] from the RNTN output. Then, the new sentiment scores of tweets A and B are both 0.1.

Using these scores, we can easily understand the overall sentiment and compare the sentiments of different tweets. However, this sentiment score does not show a more detailed view of how polarized the sentiment trends are. Both of the example tweets above have 0.1 as their sentiment scores, but these two have a substantial difference in sentiment distribution. Tweet A is more neutral, while tweet B is more polarized, which is hidden in sentiment scores. Therefore, from sentiment scores, we classify each tweet into positive if the score is greater than 0.33, negative if less than -0.33 , and neutral if between -0.33 and 0.33 . This process allows us to not only identify more distinct extreme sentiments while not excluding neutral sentiment but also understand the polarization of each tweet based on the distribution of its sentiment classes. Also, we can easily compare the distribution of three sentiment classes over time, while sentiment scores help us capture overall changes in sentiment trends and identify similar trends across geography.

3.3. Spatial time series analysis

3.3.1. Identifying spatio-topical sentiment trends

To identify the spatio-topical variation of sentiment trends, we first construct time series of sentiment scores for each geographic area and subtopic. We calculate the average sentiment score of all tweets in each area and for each topic. However, partitioning tweets by geographic units and subtopics makes the time series data even more sparse, causing spurious fluctuations in the sentiment score. To address this issue, we employ time series smoothing and fine-tuning of temporal and spatial dimensions. First, we apply a moving average window for smoothing the time series of sentiment scores. For example, a three-day moving window smooths the sentiment score for a day by calculating the average sentiment of the day of estimation, the day before and the day after. Our objective is to smooth the data as little as possible to remove large fluctuations and keep the trends as similar as possible to the original time series. Second, we apply a fine-tuning approach to reaggregate data both temporally and spatially. We employ varying temporal periods to determine and summarize sentiment scores in temporal units to adjust for bursty time series distribution. Also, we further regroup geographic units with sparse time series data based on their trend similarity and geopolitical coherence. We explain the fine-tuning of the temporal and spatial aggregation with our case study in [section 4.1](#).

3.3.2. Grouping places with similar spatio-topical sentiment trends

After extracting spatio-topical sentiment trends, we distinguish these trends among different geographies. We first identify the groups of geographic areas that share similar sentiment trends by evaluating the similarity of the sentiment trends between geographic areas for each subtopic by using dynamic time warping (DTW). First, to measure the similarity between two sentiment time series, we use DTW which compares the patterns of two or more time series and identifies non-linear relations between them by handling different lengths, noise, shifts, and amplitude changes (Brown & Rabiner, 1982; Stübinger & Schneider, 2020). DTW computes the minimum distance of an optimal match between two time series, so the DTW distance between two time series becomes smaller as they have more similar trends. We compute the DTW distance for every pair of two sentiment trends (i.e. sentiment trends of two geographic areas) and calculate the pairwise similarity among all areas. We create an undirected network graph of geographic areas based on their similarities. In this graph, a node represents a geographic area (e.g. state), and an edge represents the time series similarity of two geographic areas, and the edge weight is determined by the DTW distance. Because we initially calculate DTW for edges between all nodes, a large proportion of edges has large DTW distances showing little similarity between nodes. Therefore, we prune the edges by eliminating the pairs whose similarities are less than the upper quartile of all similarities while keeping at least one edge for every node. We then perform the Louvain community detection method (Blondel et al., 2008) to identify nodes (geographic areas) that share similar sentiment trends for each topic. We later use the resulting clusters of the Louvain community detection for visualizing similar spatio-topical trends using a time series cartogram.

3.4. Geovisualization

To visualize the spatio-topical sentiment trends, we develop a non-contiguous rectangular cartogram for each topic. Each geographic unit is represented with a bar chart that illustrates the distribution of sentiment classes – positive, neutral, and negative – over time. In these bar charts, the sum of positive, neutral, and negative classes is 100%. So, a bar chart with 10% positive and 10% negative indicates that it also has 80% neutral sentiment. Therefore, the charts also include neutral tweets, which are illustrated by the empty spaces up and down from the positive and negative percentage values. We shrank the charts to keep them compact

and highlight the extreme positive and negative tweet distributions. Therefore, the white spaces above and below the positive and negative bars are shorter and end around approximately 35% rather than 50%. However, one could still compare the relative presence of neutral tweets for each topic and each geographic unit by comparing the white spaces in each chart.

Along with charts for each geographic unit (i.e. state in our case study), we include a global time series chart (i.e. “All States” in our case study) to represent the average sentiment distribution per subtopic. We differentiate chart sizes based on the number of tweets belonged to each geographic unit for each topic. To determine the size of each chart for each geographic unit, one can use proportional scaling based on the number of tweets. However, using proportional scaling of charts generate vast differences between the chart sizes. Thus, we employ Jenks natural breaks classification to group charts into three sizes. Using classification to differentiate chart sizes helps readers perceive the size differences between the charts and still be able to distinguish sentiment trends between charts. The placement of these charts is determined based on the size of each chart and the location of each geographic unit.

From these charts, map readers could observe sentiment trends by comparing bar charts; however, identifying similar trends is challenging. To help alleviate the comparison task, we color chart frames and grid lines that reference timeframes on the x-axis and sentiment scores on the y-axis in each chart using a distinct color hue based on the geographic clusters sharing sentiment trends. Also, we include a chart with time series graphs of sentiment score by cluster on the bottom left to make it easy to observe how sentiment trends change for each cluster. A word cloud representing each topic is also included on the bottom right to explain the semantics of each topic of the sentiment trends.

4. Results

4.1. Case study and data processing

We use the 2016 United States presidential debates as a case study. The first debate took place on 26 September 2016, followed by the second debate on 9 October 2016, and the third debate on 19 October 2016. The participants were two major presidential candidates, the Democratic nominee Hillary Clinton, and the Republican nominee Donald Trump. The debates sparked active discussions on social media, particularly Twitter, which makes it interesting to investigate how sentiment trends varied across different

geographies and over different subtopics throughout the progression of the debates.

Presidential debates have been a focal point of considerable scholarship (Kraus, 2013). Of this work, our study speaks most to the literature on how presidential debates are discussed on social media platforms like Twitter (Zheng & Shahin, 2020). Since debate-watching produces more informed citizens and makes citizens more likely to participate in politics (Jamieson & Birdsell, 1990), looking at how presidential debates are discussed online is especially important to understanding American democracy (Houston et al., 2013). This is especially true with regard to the 2016 presidential debates (Jennings et al., 2020), which were viewed by 84 million people (Grynbaum, 2016) and was the most tweeted presidential debate at that time (White, 2016). Although previous scholars have discussed the importance of live-tweeting this debate and others (Houston et al., 2013), the present study considers how online communities can form around such discussions. Instead of thinking of Twitter users as a collection of individuals, we are interested in how Twitter users organize themselves – through their shared discourse – into organic clusters, many of which transcend traditional geographic boundaries.

To understand this collective discourse, we first collect Twitter data based on the keyword in tweet texts and hashtags. For the keyword, we explicitly use only “debates” to study conversational discourse related to the debate events. In this study, the keyword “debates” can represent the presidential debates of 2016 because we study a limited time period during these debates that are major social events, as well as leading topics on Twitter. Indeed, “debates” is the main hashtag used during the three presidential debates. Using “debates” as a search keyword also covers other related keywords such as presidential debates, 2016 debates, election debates, etc. This allows us to avoid bias that would be caused by varying distributions of other potential keywords, such as the names of presidential candidates. Therefore, using the keyword “debates”, we collect about 5 million tweets during our study period, from 26 September 2016, to 25 October 2016 (Figure 3). Among the collected tweets, we filter geo-located original tweets by excluding retweets, duplicates, tweets by bots, and tweets without any geotags or place tags. After filtering out those tweets, we finally obtain 825,712 tweets which we use for the analysis in this study. We use states as the geographic unit of analysis for two reasons. First, our data collection is keyword-based with a large proportion of tweets geolocated at the state level. Second, state-level aggregation allows us to reduce data sparsity especially because the time series

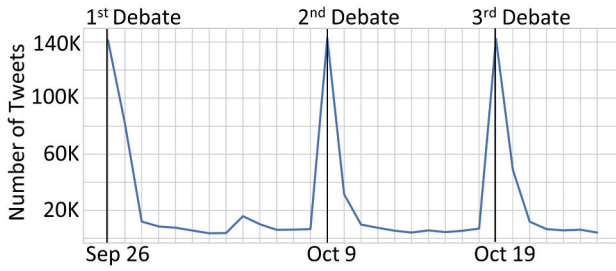


Figure 3. Temporal changes in daily tweet frequency during three debates.

data are divided by geography and subtopics. In a different dataset and case study, one may employ finer geographic units such as counties to employ our methodology and identify geopolitical regions of similar sentiment trends that may or may not follow the administrative boundaries of states.

Figure 3 shows that the number of tweets increases dramatically when each debate starts and decreases exponentially right after the debate ends. Here, the reason that the maximum number of tweets for each debate is similar is due to the limitation of the Twitter Streaming API that returns only up to 1% of the total tweets. Because the data size varies over time, we first apply the fine-tuning approach to define the temporal unit of aggregation to avoid the data sparsity issue of some units. Considering that trends of the data will also be further split by topic and geography dimensions, the effect of data sparsity will even be greater a day after each debate. If we use a fixed-length periodization such as a day-long or 12-hour-long period, data sparsity will cause sentiment trends with fewer tweets to have spurious variations. Therefore, we use varying length periods to accommodate two objectives: (1) allocating approximately the same or a similar number of tweets to each period (e.g. similar to quantile classification) (2) considering the timeline of the debates to determine the breaks during and shortly after each debate. Finally, the varying lengths of time periods we use are 30 min from each debate’s starting time, 1 hour after the first 30 min, then one and a half hours, then 6 hours, then a day, and then the rest of the time before the starting time of the next debate. Since the format of the first and the third debates has six segments with 15 min each, we define the time length not to break in the middle of each segment and to have a similar number of tweets without strong fluctuations within each period.

We limit the study area to the 48 states in the contiguous United States. We exclude nine states: Montana, Idaho, North Dakota, South Dakota, Wyoming, Arkansas, West Virginia, Rhode Island, and Vermont, as these states do not have sufficient tweets to employ

the time series analysis. Also, we group 27 states with sparse time series data into seven regions based on known socio-economic and cultural divisions in human geography. These regions are the rest of Northwest, Northeast, Mid-Atlantic, Southeast, Southwest, Lower Midwest, and Upper Midwest (see [Appendix A](#)). As a result, there is a total of 20 geographic units formed by 13 states and 7 regions.

4.2. Topic modeling, clustering, and evolution

To capture subtopics emerging and disappearing in a short period, we partition the temporal extent of this study into three periods based on the timeline of the three debates. Each period starts at the debate starting time and ends at the next debate starting time. For example, the first time period is from September 26 at 9 p.m. to October 9 at 9 p.m. (Eastern Standard Time). As described in [Section 3.2.1](#), we perform topic modeling to identify the optimal topic model and the associated topics for each period. We then cluster all topics of the three optimal topic models by performing the Louvain community detection on the topic similarity network, which is constructed based on the word-topic probabilities of topics in all three models. Here we use the term “subtopic” to distinguish a topic of each model from a topic cluster, which we refer to as “topic.” Grouping subtopics within and across different temporal topic models is essential because tracing temporal trends requires robust time series sentiment scores. Without grouping, most emerging topics suffer from sparse time series data, and therefore, it becomes difficult to identify their temporal trends.

As a result of grouping subtopics, we identify six major topics which are illustrated in [Figure 4](#) as word clouds. The words in each word cloud are the main words appearing in that topic, and the size of each word is proportional to the frequency of that word within each topic. At first glance, for example, Topic 0 seems to be related to Donald Trump. However, upon closer examination, the topic has more to do with the candidates themselves, as demonstrated by the size of “trump” and “hillary” in [Figure 4](#) word cloud. Given that the frequency of the former is larger than the latter, discussions of Donald Trump seem to be much more influential on this topic than discussions of Hillary Clinton, which is consistent with the broader narrative of the 2016 presidential election (Sides et al., 2017). On the other hand, Topic 1 seems to focus on Chris Wallace and the debates in general. Although the word “debates” is the most prevalent in this topic, the prominence of “chris” and “wallace” is most likely in relation to the performance of Chris Wallace as a moderator during

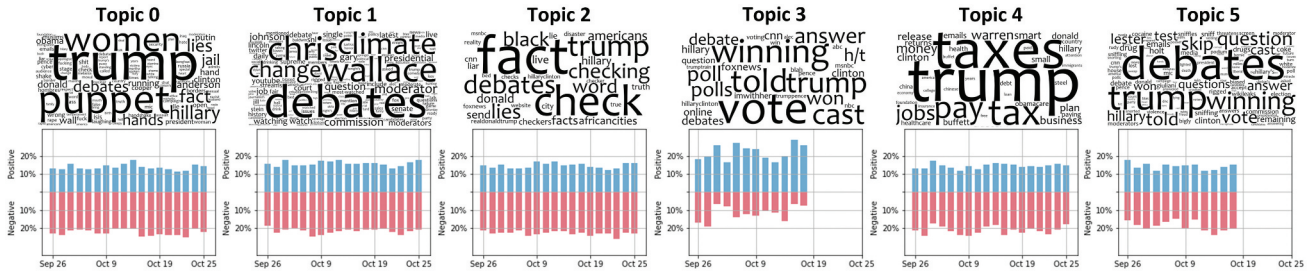


Figure 4. Word cloud and sentiment trends of the six topics.

the third presidential debate. Not only was he lauded for his ability to produce a productive dialogue between the candidates (Byers, 2016), but many thought he was able to “tame” Donald Trump, which is one of the reasons why he is the only 2016 moderator to be invited back to moderate a 2020 presidential debate. Also, in Topic 1, the word “climate” is also prominent, which may be due to the criticism about the climate change issue that is never discussed in any of the three debates. Although the other main topics are undoubtedly interesting, topics 0 (a.k.a., “trump”) and 1 (a.k.a., “debates”) are prevalent enough for us to conduct the analysis and visualization of spatio-topical sentiment trends that is the focal point of this study.

Although grouping subtopics into major topics seems to lose the variation of subtopics within each topic, one can always break each topic to its original subtopics and trace their semantic and temporal patterns. Figure 5 illustrates the words that form subtopics and their associated time series trend of sentiment scores. In Figure 5, Topic 0 consists of Subtopic 1, 4, and 8 of the first period, Subtopic 7, 10, 16, and 22 of the second period, and Subtopic 2, 4, and 11 of the third period. Although Topic 0 is overall about two candidates, there exist different subtopics, such as Trump’s claim about Obama and Clinton being founders of ISIS

(Gibson & Holland, 2016) as appeared in Subtopic 1 of Period 1.

4.3. Global sentiment trends by subtopics

To examine the sentiment trends, we compute the percentage of each of positive, neutral, and negative tweets for each topic in each time frame, which are shown as bar charts in Figure 4. As discussed in Section 3.4, these bar charts illustrate the neutral sentiment using the empty space. For example, if a bar chart has 20% of positive tweets and 10% of negative tweets, it means that 70% of tweets are neutral. While positive tweets are almost always below 20% of all tweets for all topics except Topic 3, negative tweets correspond to approximately 20% of all tweets in most topics. Therefore, nearly 50% to 60% of tweets are neutral for all topics. Using these bar charts, Figure 4 illustrates global time series trends for each topic and highlights how the sentiment trend of each topic evolves over time. For example, Topic 3 and Topic 5 appear only before the day of the third debate. Since we extract subtopics separately for three different time periods (i.e. each period starts at each debate starting time and ends at the next debate starting time), this trend shows that subtopics in Topic 3, which is about the presidential polls, and Topic 5, which is about the fact-checking of

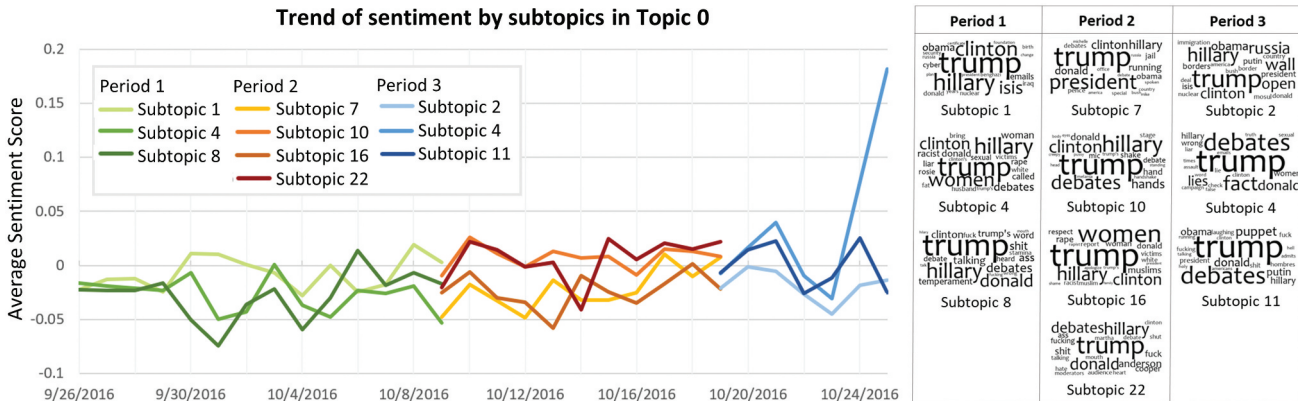


Figure 5. Evolution of subtopics in Topic 0.

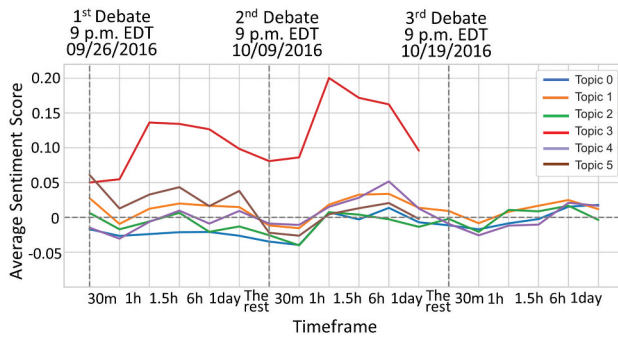


Figure 6. Time series of average sentiment score by topic.

claims of the candidates, have been actively discussed only during the first two debate periods.

Although Figure 4 is useful for comparing relative amounts of the three sentiment classifications over time, we illustrate the average sentiment score of each topic in Figure 6 to understand the overall changes of sentiment trend by topic. On the y-axis, scores above 0 are positive and below 0 are negative. The trends in Figure 6 highlight fluctuations between positive and negative sentiments over time only except Topic 3. Topic 3, which is about the presidential election polls, has positive sentiment over time. Besides Topic 3, the other five topics have only small differences in their patterns. For example, Topic 0, which is related to two candidates, shows overall negative sentiment rather than positive, while Topic 1, which is about the debates in general, exhibits positive to neutral sentiment.

4.3. Spatio-topical sentiment trends

We further partition the global sentiment trends by topics and states. We compare the sentiment trends of all states to identify which geographic areas share similar sentiment trends for a given topic. Following the methodology described in Section 3.3.2, we compute the DTW distance to measure the similarity between

sentiment trends. As the two states have more similar sentiment trends, DTW distance becomes smaller. We create a network graph where our 20 geographic units (i.e. 13 states and 7 regions) are nodes, and similarities between sentiment trends of every pair of two units are edges. To calculate the edge weights, we reverse the DTW distance values since DTW distance is larger for dissimilar sentiment trends. Then, we conduct the Louvain community detection method to group the states and regions into geographic clusters that share similar trends. Note that these geographic clusters may be formed by geographically disjoint states because the Louvain method does not enforce spatial contiguity in deriving communities.

Figure 7 shows the geographic clusters identified for Topic 0 and Topic 1. States and regions sharing the same color are in the same geographic cluster, and the order of clusters has no meaning. This figure shows that adjacent states do not always share a similar sentiment trend, and the geographic clusters that share similar sentiment trends also vary across topics. For example, in Topic 0, Florida is in the same cluster as Texas, whereas in Topic 1, these states are in different clusters. The same could be said for Texas and Georgia. These two states are in different clusters in Topic 0 but in the same cluster in Topic 1. Traditionally, scholars would think of Florida, Texas, and Georgia as being part of the deep South and a relatively homogenous voting bloc. However, Figure 7 shows that these states often comprise different communities online, depending on the topic. This does not mean that these states are not politically aligned, but rather, it suggests that traditional notions of homogeneity and community may need to be reconsidered when considering social media platforms, like Twitter.

To take a deeper look into spatio-topical sentiment trends, we visualize the sentiment distribution of each state using time series cartogram (Figure 8). Each state or region is represented with a time series bar chart

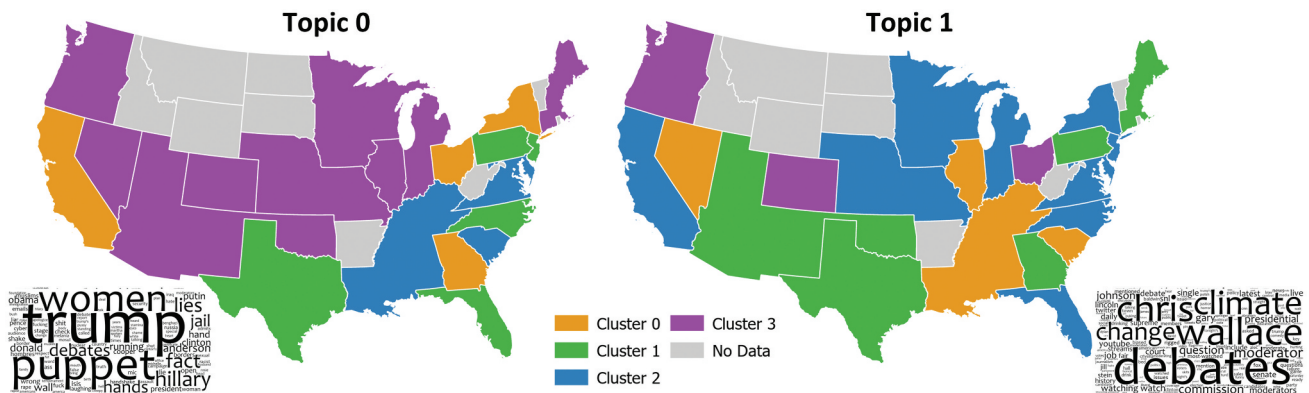
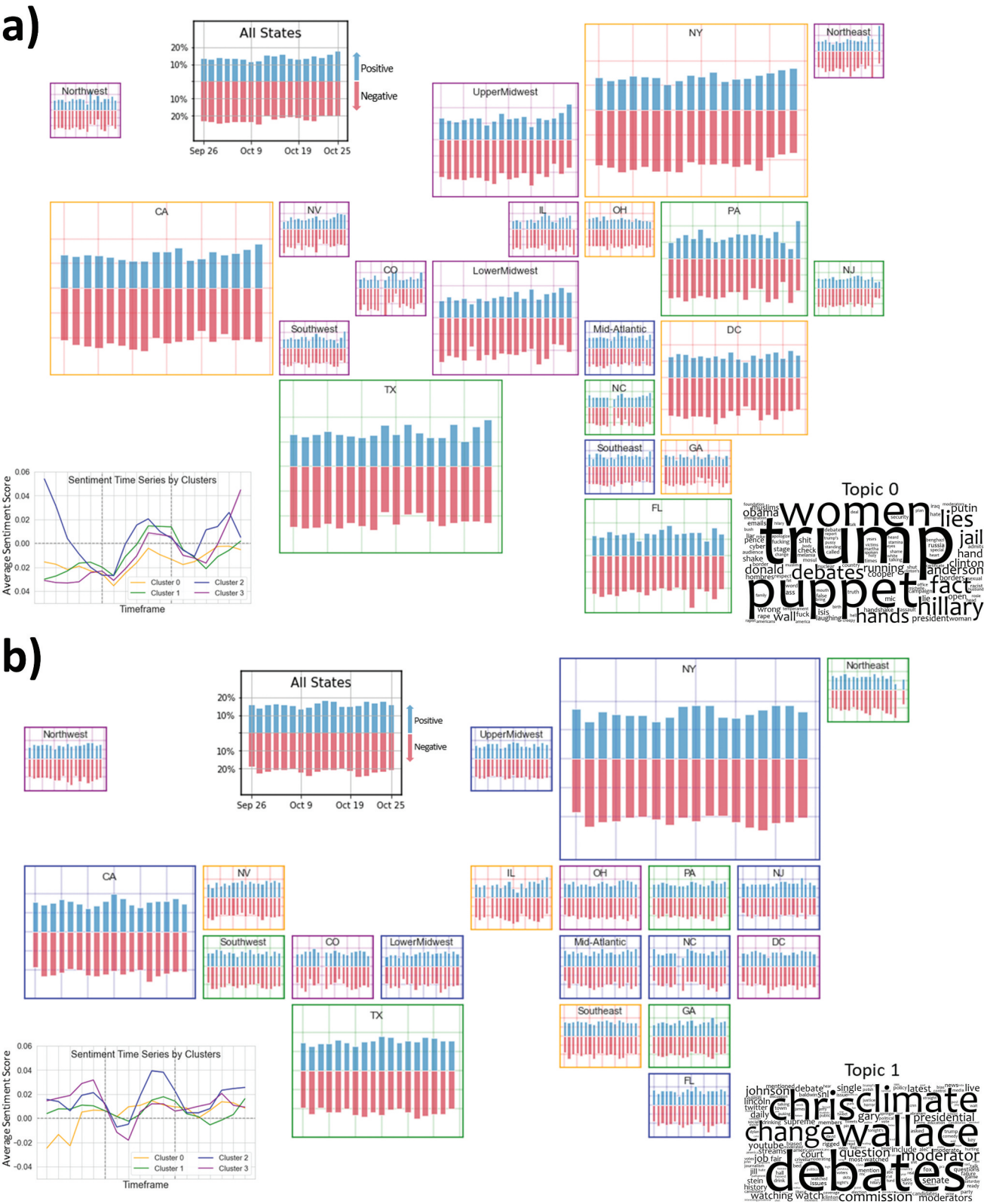


Figure 7. Clusters with high similarities of sentiment time series for Topic 0 (left) and Topic 1 (right).



illustrating the sentiment trend. Chart sizes are determined based on the number of tweets that belong to the topic of attention across all geographic units. We use the geographical clusters in Figure 7 to visually encode each state or region based on their clusters. We do this by color-coding the frame and the grid lines of each bar chart depicted in the spatio-topical sentiment cartograms (Figure 8). On the bottom left of Figure 8, there is a line graph that illustrates the sentiment time series of each geographic cluster. We include this to help understand the changes in general sentiment trends by clusters, while bar charts allow us to compare the relative proportion of negative, positive, and neutral sentiments. The word cloud for each topic is also included on the bottom right to explain what each topic is about.

In Topic 0 (Figure 8a), which is about the candidates, Texas, Pennsylvania, and Florida are in the same cluster in green, while New York, California, and DC are in the orange cluster. As we can see from the sentiment time series chart on the lower left, states in the green cluster have negative sentiment mostly during the first and third debates but the trend turns positive during the second debate. On the other hand, states in the orange cluster keep negative sentiment over all periods despite some fluctuations. In Topic 1 (Figure 8b), which is a discussion about the debates and the moderator, the sentiment time series chart reveals that Topic 1 is overall more positive than Topic 0 in all geographic clusters. As we discussed in the previous Section 4.3, we often think southern states, like Georgia, Florida, and Texas, are politically interchangeable, ultimately forming a common voting bloc, but, at least online, this is not the case. In Topic 0, Florida and Texas are in the same cluster, whereas in Topic 1, they are in different clusters, and instead, Texas and Georgia are in the same cluster. Therefore, Figure 8 reveals how online communities can (and often do) transcend traditional political boundaries, as Figure 7 also shows, but with more detailed information about sentiment distributions over time.

5. Discussion and conclusion

This study introduces an approach that integrates natural language processing with spatial time series analysis and geovisualization to identify spatio-topical sentiment trends on Twitter. Using the 2016 presidential debates as a case study, the results show that temporal distributions of sentiments vary across different subtopics and geographies. Our findings also reveal that adjacent states do not always share similar sentiment trends, and the geographic clusters with similar sentiment trends also vary across topics. Failing to consider these variations may result in misunderstanding public

discourse and sentiments that are diverse and spatio-temporally dynamic. For example, we find that states traditionally associated with one another politically often differ in terms of their online discourse. This suggests that our approach can not only help better explain key spatiotemporal variations on Twitter but may also help scholars gather new insights into how online communities form and evolve during discussions of major political events.

While the core focus of our article revolves around the 2016 presidential debates to contextualize our findings with the series of events and developments during the debates, we have also broadened our scope by including a case study on tweets during hurricane Irma in 2017. The results of this Irma case study align well with the major findings of our study that the patterns of sentiment changes over time vary substantially by different topics and across geographies, even at a smaller scale than state. The Irma case study reinforces the findings presented in the presidential debates case study, highlighting the dynamic nature of temporal sentiment patterns across different topics and geographic regions. Furthermore, it underscores that the geographic clusters exhibiting similar sentiment patterns can vary across topics and may not necessarily be adjacent to one another. This supplementary investigation serves as a compelling demonstration of the broad generalizability and practical applicability of our workflow across diverse research domains.

There are, however, some limitations in this study. First, although grouping data spatially and temporally is inevitable to overcome the sparsity problem of the time series data, the way how to group the data may affect the results. For example, when we aggregate states, we consider spatial adjacency in addition to cultural coherence, but it would also be possible to use the similarity of sentiment time series instead of the adjacency to group states. To make the results more reliable, evaluations on how different data grouping methods change the results are needed in the future work. Also, we only included tweets written in English in the analysis, which may result in unintentionally excluding the users whose first language is not English. This study, particularly, uses political events as a case study, so people's reactions to those events may significantly vary across different ethnicities. So, our results may include a bias due to excluding non-English tweets, which we need further investigation.

Despite these limitations, our approach provides an important foundation for future work. For example, online discussions are also used to assess public health trends (Paul et al., 2014). If scholars were able to look simultaneously at spatial and topical dimensions of

sentiment trends within these discussions, then it would make assessing public health trends more precise and make interventions more effective. The same could be said for other types of forecasting (Nisar & Yeung, 2018). Whether it is an upcoming election (Tumasjan et al., 2011) or general demographic trends (McCormick et al., 2017), online discussions may prove more useful when the underlying dynamics are modeled simultaneously. Regardless of the application, what we show in this paper provides a blueprint for these future endeavors. Although online communities are often tied to physical places, sometimes they transcend these boundaries, especially when discussing important political events. By modeling these dynamics, we provide scholars with the tools necessary to begin exploring similar relationships in their own data. We look forward to seeing future efforts and how they help shape our understanding of communities on- and offline.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable insights. The reviewers' constructive comments greatly contributed to the improvement of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Hoeyun Kwon  <http://orcid.org/0000-0001-6571-2170>
Caglar Koylu  <http://orcid.org/0000-0001-6619-6366>
Bryce J. Dietrich  <http://orcid.org/0000-0002-9781-3088>

Data availability statement

Twitter data used in this study are openly available in figshare at <https://doi.org/10.6084/m9.figshare.20277840.v1>. The shared data contain tweet IDs related to a series of three presidential debates in 2016 between the dates of September 26 and 26 October 2016.

References

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022. <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory & Experiment*, 2008(10), 10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- Brown, M., & Rabiner, L. (1982). Dynamic time warping for isolated word recognition based on ordered graph searching techniques. *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 7, 1255–1258. <https://doi.org/10.1109/ICASSP.1982.1171695>
- Byers, D. (2016, October 19). Chris Wallace delivers sterling performance as debate moderator. *CNNMoney*. <https://money.cnn.com/2016/10/19/media/chris-wallace-moderator-presidential-debate/index.html>
- Carmines, E. G., Ensley, M. J., & Wagner, M. W. (2016). Ideological heterogeneity and the rise of Donald Trump. *The Forum*, 14(4), 385–397. <https://doi.org/10.1515/forum-2016-0036>
- Diakopoulos, N. A., & Shamma, D. A. (2010). Characterizing debate performance via aggregated twitter sentiment. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1195–1198. <https://doi.org/10.1145/1753326.1753504>
- Dunkel, A., Andrienko, G., Andrienko, N., Burghardt, D., Hauthal, E., & Purves, R. (2019). A conceptual framework for studying collective reactions to events in location-based social media. *International Journal of Geographical Information Science*, 33(4), 780–804. <https://doi.org/10.1080/13658816.2018.1546390>
- Du, J., Xu, J., Song, H.-Y., & Tao, C. (2017). Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. *BMC Medical Informatics and Decision Making*, 17(2), 69. <https://doi.org/10.1186/s12911-017-0469-6>
- Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., Weeg, C., Larson, E. E., Ungar, L. H., & Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26(2), 159–169. <https://doi.org/10.1177/0956797614557867>
- Garske, S. I., Elayan, S., Sykora, M., Edry, T., Grabenhenrich, L. B., Galea, S., Lowe, S. R., & Gruebner, O. (2021). Space-time dependence of emotions on Twitter after a natural disaster. *International Journal of Environmental Research and Public Health*, 18(10), 5292. <https://doi.org/10.3390/ijerph18105292>
- Gibson, G., & Holland, S. (2016, August 11). Trump calls Obama, Clinton Islamic State “co-founders,” draws rebuke. *Reuters*. <https://www.reuters.com/article/us-usa-election-trump-idUSKCN10M146>
- Gruebner, O., Lowe, S. R., Sykora, M., Shankardass, K., Subramanian, S. V., & Galea, S. (2018). Spatio-temporal distribution of negative emotions in New York City after a natural disaster as seen in social media. *International Journal of Environmental Research and Public Health*, 15(10), 2275. Article 10. <https://doi.org/10.3390/ijerph15102275>
- Grynbaum, M. M. (2016, September 27). At 84 million viewers, debate was the most-watched ever. *The New York Times*. <https://www.nytimes.com/2016/09/28/business/media/at-nearly-84-million-viewers-debate-may-be-the-most-watched-ever.html>
- Han, X., Wang, J., Zhang, M., & Wang, X. (2020). Using social media to mine and analyze public opinion related to COVID-19 in China. *International Journal of Environmental Research and Public Health*, 17(8), 2788. Article 8. <https://doi.org/10.3390/ijerph17082788>

- Houston, J. B., Hawthorne, J., Spialek, M. L., Greenwood, M., & McKinney, M. S. (2013). Tweeting during presidential debates: Effect on candidate evaluations and debate attitudes. *Argumentation & Advocacy*, 49(4), 301–311. <https://doi.org/10.1080/00028533.2013.11821804>
- Houston, J. B., McKinney, M. S., Hawthorne, J., & Spialek, M. L. (2013). Frequency of tweeting during presidential debates: Effect on debate attitudes and knowledge. *Communication Studies*, 64(5), 548–560. <https://doi.org/10.1080/10510974.2013.832693>
- Huang, A. (2008). Similarity measures for text document clustering. *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, Christchurch, New Zealand, 4, 9–56. <https://www.academia.edu/download/44422710/SMTP.pdf>
- Hu, T., Wang, S., Luo, W., Zhang, M., Huang, X., Yan, Y., Liu, R., Ly, K., Kacker, V., She, B., & Li, Z. (2021). Revealing public opinion towards COVID-19 vaccines with Twitter data in the United States: A spatiotemporal perspective. *MedRxiv*. <https://doi.org/10.1101/2021.06.02.21258233>
- Jamieson, K. H., & Birdsell, D. S. (1990). *Presidential debates: The challenge of creating an informed electorate*. Oxford University Press.
- Janowicz, K., McKenzie, G., Hu, Y., Zhu, R., & Gao, S. (2019). Using semantic signatures for social sensing in urban environments. In C. Antoniou, L. Dimitriou, & F. Pereira (Eds.), *Mobility patterns, big data and transport analytics* (pp. 31–54). Elsevier. <https://doi.org/10.1016/B978-0-12-812970-8.00003-8>
- Jennings, F. J., Warner, B. R., McKinney, M. S., Kearney, C. C., Funk, M. E., & Bramlett, J. C. (2020). Learning from presidential debates: Who learns the most and why? *Communication Studies*, 71(5), 896–910. <https://doi.org/10.1080/10510974.2020.1807377>
- Jeong, B., Yoon, J., & Lee, J.-M. (2019). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management*, 48, 280–290. <https://doi.org/10.1016/j.ijinfomgt.2017.09.009>
- Key, V. O., & Heard, A. (1949). *Southern politics in state and nation*. AA Knopf. <https://doi.org/10.1002/ncr.4110390714>
- Kinsella, C., McTague, C., & Raleigh, K. N. (2015). Unmasking geographic polarization and clustering: A micro-scalar analysis of partisan voting behavior. *Applied Geography*, 62, 404–419. <https://doi.org/10.1016/j.apgeog.2015.04.022>
- Koylu, C. (2018). Uncovering geo-social semantics from the Twitter mention network: An integrated approach using spatial network smoothing and topic modeling. In S. L. Shaw & D. Sui (Eds.), *Human dynamics research in smart and connected communities*. (pp. 163–179). Springer. https://doi.org/10.1007/978-3-319-73247-3_9
- Koylu, C. (2019). Modeling and visualizing semantic and spatio-temporal evolution of topics in interpersonal communication on Twitter. *International Journal of Geographical Information Science*, 33(4), 805–832. <https://doi.org/10.1080/13658816.2018.1458987>
- Koylu, C., Larson, R., Dietrich, B. J., & Lee, K.-P. (2019). CarSenToGram: Geovisual text analytics for exploring spatiotemporal variation in public discourse on Twitter. *Cartography and Geographic Information Science*, 46(1), 57–71. <https://doi.org/10.1080/15230406.2018.1510343>
- Kraus, S. (2013). *Televised presidential debates and public policy*. Routledge. <https://doi.org/10.4324/9781315044859>
- Li, J., Chen, S., Chen, W., Andrienko, G., & Andrienko, N. (2020). Semantics-space-time cube: A conceptual framework for systematic analysis of texts in space and time. *IEEE Transactions on Visualization and Computer Graphics*, 26(4), 1789–1806. <https://doi.org/10.1109/TVCG.2018.2882449>
- Liu, R., Yao, X., Guo, C., & Wei, X. (2021). Can we forecast presidential election using Twitter data? An integrative modelling approach. *Annals of GIS*, 27(1), 43–56. <https://doi.org/10.1080/19475683.2020.1829704>
- Martin, Y., Li, Z., Ge, Y., & Huang, X. (2021). Introducing Twitter daily estimates of residents and non-residents at the county level. *Social Sciences*, 10(6), 227. Article 6. <https://doi.org/10.3390/socsci10060227>
- McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2017). Using Twitter for demographic and social Science Research: Tools for data collection and processing. *Sociological Methods & Research*, 46(3), 390–421. <https://doi.org/10.1177/0049124115605339>
- Nisar, T. M., & Yeung, M. (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. *The Journal of Finance and Data Science*, 4(2), 101–119. <https://doi.org/10.1016/j.jfds.2017.11.002>
- Parimala, M., Priya, R. M. S., Reddy, M. P. K., Chowdhary, C. L., Poluru, R. K., & Khan, S. (2021). Spatiotemporal-based sentiment analysis on tweets for risk assessment of event using deep learning approach. *Software: Practice and Experience*, 51(3), 550–570. <https://doi.org/10.1002/spe.2851>
- Paul, M. J., Dredze, M., & Broniatowski, D. (2014). Twitter improves influenza forecasting. *PLoS Currents*, 6. <https://doi.org/10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117>
- Robertson, C. T., Dutton, W. H., Ackland, R., & Peng, T.-Q. (2019). The democratic role of social media in political debates: The use of Twitter in the first televised US presidential debate of 2016. *Journal of Information Technology & Politics*, 16(2), 105–118. <https://doi.org/10.1080/19331681.2019.1590283>
- Sides, J., Tesler, M., & Vavreck, L. (2017). The 2016 U.S. Election: How Trump lost and won. *Journal of Democracy*, 28(2), 34–44. <https://doi.org/10.1353/jod.2017.0022>
- Sit, M. A., Koylu, C., & Demir, I. (2019). Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: A case study of hurricane Irma. *International Journal of Digital Earth*, 12(11), 1205–1229. <https://doi.org/10.1080/17538947.2018.1563219>
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (1631–1642). <https://aclanthology.org/D13-1170.pdf>
- Stübinger, J., & Schneider, L. (2020). Epidemiology of coronavirus COVID-19: Forecasting the future incidence in different countries. *Healthcare*, 8(2), 99. Article 2. <https://doi.org/10.3390/healthcare8020099>

- Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, 29(4), 402–418. <https://doi.org/10.1177/0894439310386557>
- Wang, Z., & Ye, X. (2018). Social media analytics for natural disaster management. *International Journal of Geographical Information Science*, 32(1), 49–72. <https://doi.org/10.1080/13658816.2017.1367003>
- White, D. (2016, September 27). This was the most tweeted presidential debate ever. *Time*. <https://time.com/4508981/presidential-debate-twitter-clinton-trump/>
- Wu, F., Huang, Y., & Yuan, Z. (2017). Domain-specific sentiment classification via fusing sentiment knowledge from multiple sources. *Information Fusion*, 35, 26–37. <https://doi.org/10.1016/j.inffus.2016.09.001>
- Xu, F., & Keelj, V. (2014). Collective sentiment mining of microblogs in 24-hour stock price movement prediction. *2014 IEEE 16th Conference on Business Informatics*, (60–67). <https://doi.org/10.1109/CBI.2014.37>
- Yao, F., & Wang, Y. (2020). Domain-specific sentiment analysis for tweets during hurricanes (DSSA-H): A domain-adversarial neural-network-based approach. *Computers, Environment and Urban Systems*, 83, 101522. <https://doi.org/10.1016/j.compenvurbsys.2020.101522>
- Yaqub, U., Chun, S. A., Atluri, V., & Vaidya, J. (2017). Analysis of political discourse on twitter in the context of the 2016 US presidential elections. *Government Information Quarterly*, 34(4), 613–626. <https://doi.org/10.1016/j.giq.2017.11.001>
- Yuan, F., Li, M., & Liu, R. (2020). Understanding the evolutions of public responses using social media: Hurricane Matthew case study. *International Journal of Disaster Risk Reduction*, 51, 101798. <https://doi.org/10.1016/j.ijdrr.2020.101798>
- Zheng, P., & Shahin, S. (2020). Live tweeting live debates: How Twitter reflects and refracts the US political climate in a campaign season. *Information, Communication & Society*, 23(3), 337–357. <https://doi.org/10.1080/1369118X.2018.1503697>
- Zou, L., Lam, N. S. N., Cai, H., & Qiang, Y. (2018). Mining Twitter data for improved understanding of disaster resilience. *Annals of the American Association of Geographers*, 108(5), 1422–1441. <https://doi.org/10.1080/24694452.2017.1421897>

Appendix A. States and regions used in this study

Regions	States (number of states)
Northwest	Washington, Oregon (2)
Northeast	Maine, New Hampshire, Connecticut, Massachusetts (4)
Mid-Atlantic	Delaware, Maryland, Virginia (3)
Southeast	Alabama, South Carolina, Mississippi, Louisiana, Kentucky, Tennessee (6)
Southwest	Arizona, New Mexico, Utah, Oklahoma (4)
Lower Midwest	Missouri, Kansas, Nebraska (3)
Upper Midwest	Iowa, Indiana, Michigan, Wisconsin, Minnesota (5)
Individually analyzed states	California, Colorado, District of Columbia, Florida, Georgia, Illinois, North Carolina, New Jersey, Nevada, New York, Ohio, Pennsylvania, Texas (13)
Excluded states due to data sparsity	Montana, Idaho, North Dakota, South Dakota, Wyoming, Arkansas, West Virginia, Rhode Island, and Vermont (9)

Appendix B. A case study of Hurricane Irma

To demonstrate the generalizability and applicability of our workflow in other research domains, we conducted an additional case study using Hurricane Irma. In this appendix section, we intentionally omit reporting on the optimal parameterization of our entire workflow. Instead, our focus is to provide a concise summary of the key findings from our second case study, as we aim to streamline the length and complexity of our analysis. Hurricane Irma, a category-5 hurricane, occurred in September 2017 and caused almost one hundred deaths and severe infrastructure damage of billions of dollars, mainly in Southern Florida. We chose Hurricane Irma as our case study since social media, especially Twitter, were popular during that time and widely used by the public and relevant agencies for a variety of efforts including situational awareness, evacuation, and recovery (Sit et al., 2019). The Twitter data related to Irma were used for this case study. Details about how those tweets were collected and pre-processed can be found in the Sit et al. (2019)'s study.

We partitioned the temporal extent into three time periods based on the progress of Irma and the daily tweet frequency. The first time period is from September 4 to 7 (4 days), the second time period is from September 8 to 11 (4 days), and the third time period is from September 12 to 17 (6 days). As described in the methodology section, we identified the optimal topic model for each period and then clustered all topics of the three optimal topic models to capture subtopics emerging and disappearing in a short period. As a result, we identified five major topics, which are illustrated in Figure A-1 as wordclouds. In all five topics, ‘hurricane’ is the most frequent word, but upon closer examination, each topic has distinct subtopics. For example, Topic 0 is about sharing the situation of hurricane as well as hopes and prayers for affected people, while Topic 1 includes discussions of climate change influencing the frequency and severity of weather-related hazards. Topic 2 focuses on damage due to the hurricane, including power outages and flood. Notice that Topic 2 appears later than the other topics as the study period includes pre-period before the hurricane made its first landfall in Florida. Topic 3 is also about Irma but includes subtopics about hurricane Maria which occurred right after Irma. Topic 4 mainly includes tweets in Spanish. Among five topics, Topic 0 and Topic 1 were dominant than the other topics as Figure A-2 shows.

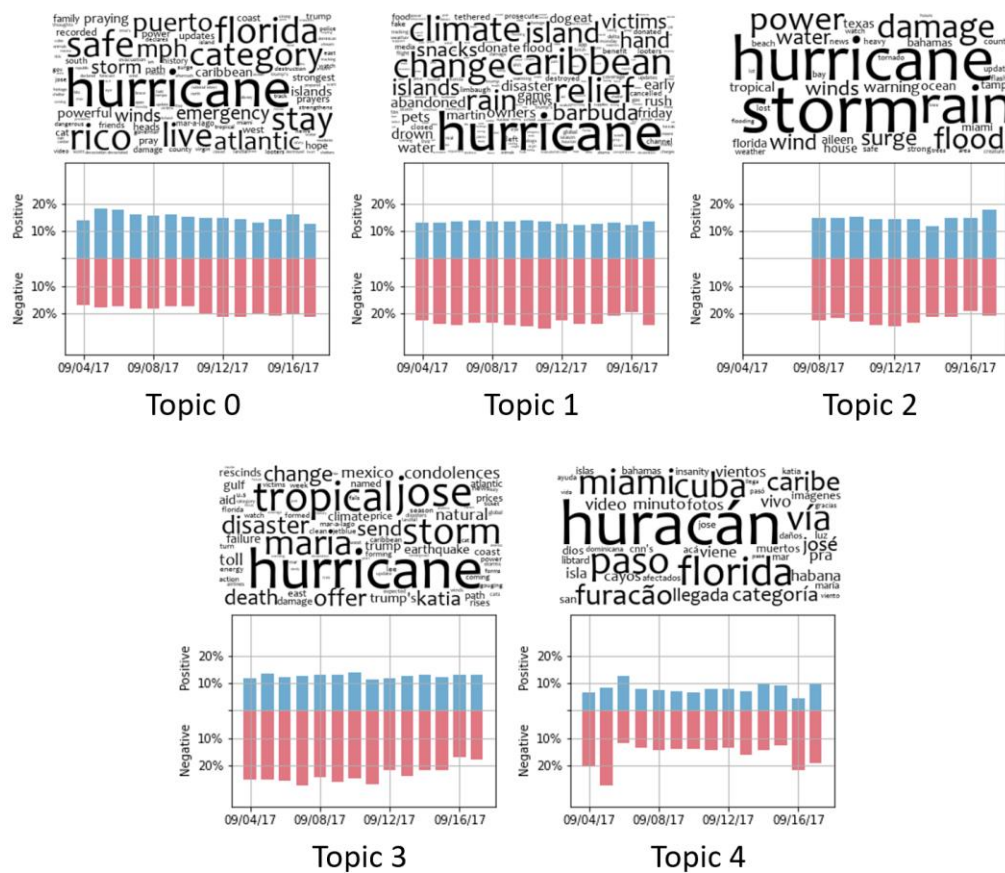


Figure A-1. Wordcloud and sentiment trends of the five topics

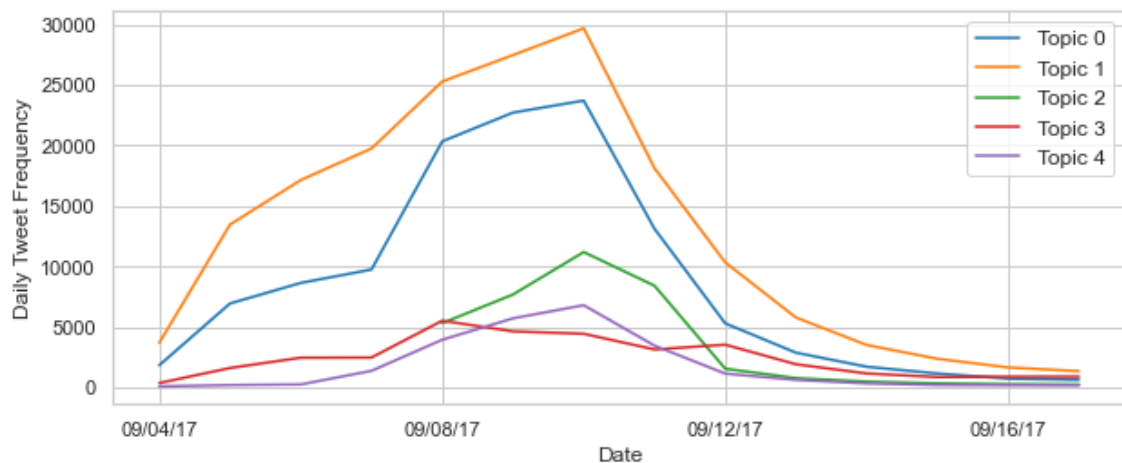


Figure A-2. Time series of daily number of tweets by topic

Figure A-3 depicts the temporal changes of sentiment by topics. Since scores above 0 are positive and below 0 are negative, most of the five topics exhibit a prevailing negative sentiment. However, it is worth noting that Topic 0 initially displayed a positive average sentiment during the first half period, which shifted to a negative sentiment. This transition can be attributed to the fact that Topic 0 includes tweets concerning prayers and hopes. It is likely that that during the initial stages of the hurricane, people tend to express more positive sentiment driven by hope. However, as time progresses and the true extent of the damage becomes increasingly evident, the sentiment gradually shifts towards negativity.

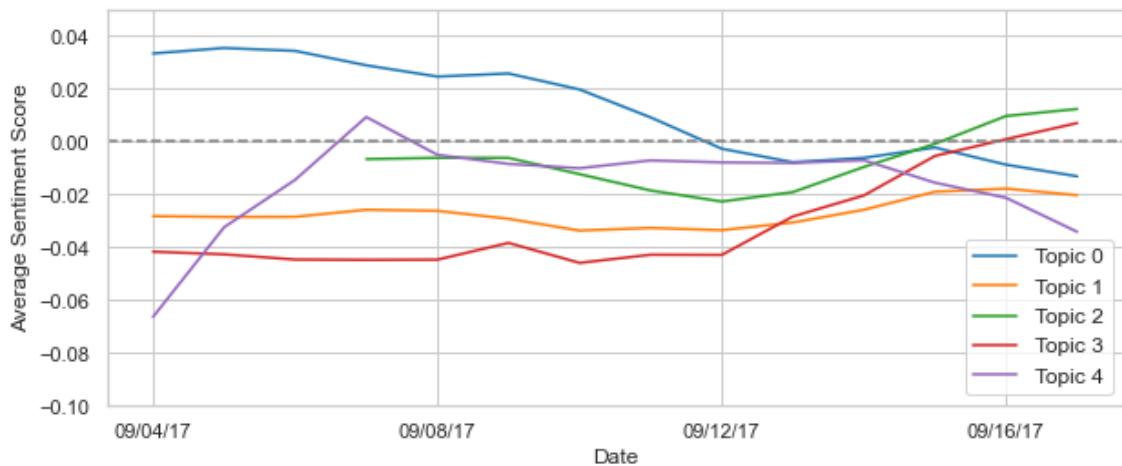


Figure A-3. Time series of average sentiment score by topic

To further understand spatio-topical sentiment trends, we performed spatial time series analysis using counties in Florida. The number of tweets varies significantly across counties in Florida, and some counties do not have sufficient data to conduct the time series analysis. To address the data sparsity issue, we first grouped counties into six regions and then investigated how sentiment trends vary across those regions and which regions share similar sentiment trends by topics. Figure A-4 illustrates six regions that we used in this case study. This grouping is based on spatial adjacency, geographic location, and data distribution.

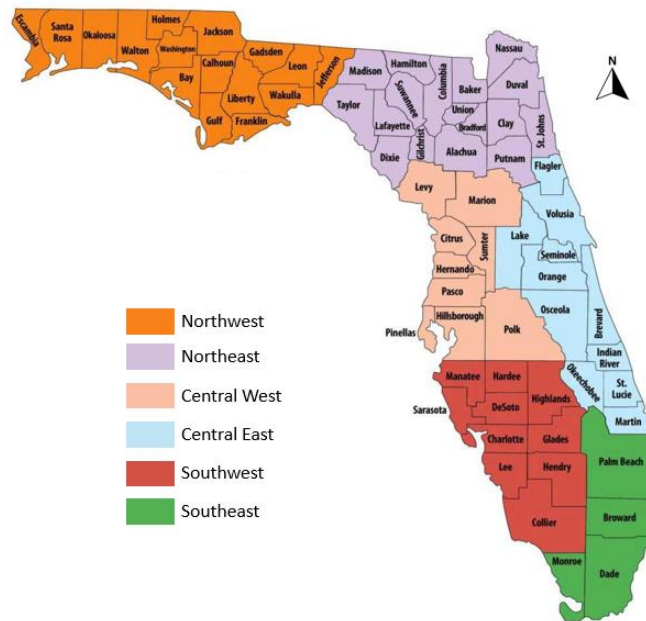


Figure A-4. Six regions grouping counties in Florida (Revised from the image by the Florida Office of Film and Entertainment <https://filminflorida.com/map-of-fl-counties/>)

Using these six regions, we identified the geographic clusters that share similar sentiment trends for each topic based on DTW distances between regions and community detection method. Next, we generated spatio-topical sentiment cartograms to represent the geographic clusters and the sentiment distribution trends of each region for different topics. These cartograms provide an illustrative depiction of both the geographic clustering and the sentiment distribution patterns across the regions. Figure A-5 shows such cartograms for two distinct topics, Topic 1 and Topic 2.

In Topic 1 (Figure A-5, upper), which is about Irma relief and climate change, all six regions have similar patterns in general, but they are classified into two clusters depending on how their sentiment trends changed at the end. Both of these clusters exhibit predominantly negative sentiments. However, it is worth noting that towards the end, the sentiment in the Northeast, Central East, and Central West regions shifted to become more positive. In contrast, the sentiment in the Southwest, Southeast, and Northwest regions displayed an intensified negativity. This observation can be

attributed to the fact that southern Florida was one of the most severely impacted areas by hurricane Irma. As a result, people residing in those areas simply shared the impact of the disaster on Twitter, leading to a predominant expression of negative sentiment.

Similarly, in Topic 2 (Figure A-5, lower), which is about damage by Irma, six regions are divided into two different clusters. Note that since this topic appears from the second time period, the bar chart of each region shows sentiment from 09/08/17, having fewer bars than charts in Topic 1 that start from 09/04/17. The two clusters in Topic 2 have similar patterns of changing sentiments, but the cluster in green has more negative sentiments than the other cluster does as the line graph in Figure A-5 demonstrates. It is not surprising that Southeast region has more negative sentiments than others since it is most severely affected. However, interestingly, Southwest and Southeast, which are in the same geographic cluster in Topic 1, are not in the same cluster in Topic 2.

These results further support the conclusions drawn in the 2016 presidential debate case study regarding the variations in temporal sentiment patterns across different topics and geographic locations. Furthermore, it underscores that the geographic clusters exhibiting similar sentiment patterns can vary across topics and may not necessarily be adjacent to one another. These observations emphasize the intricate relationship between sentiment, topics, and geography, highlighting the need for a nuanced understanding of these factors in sentiment analysis.

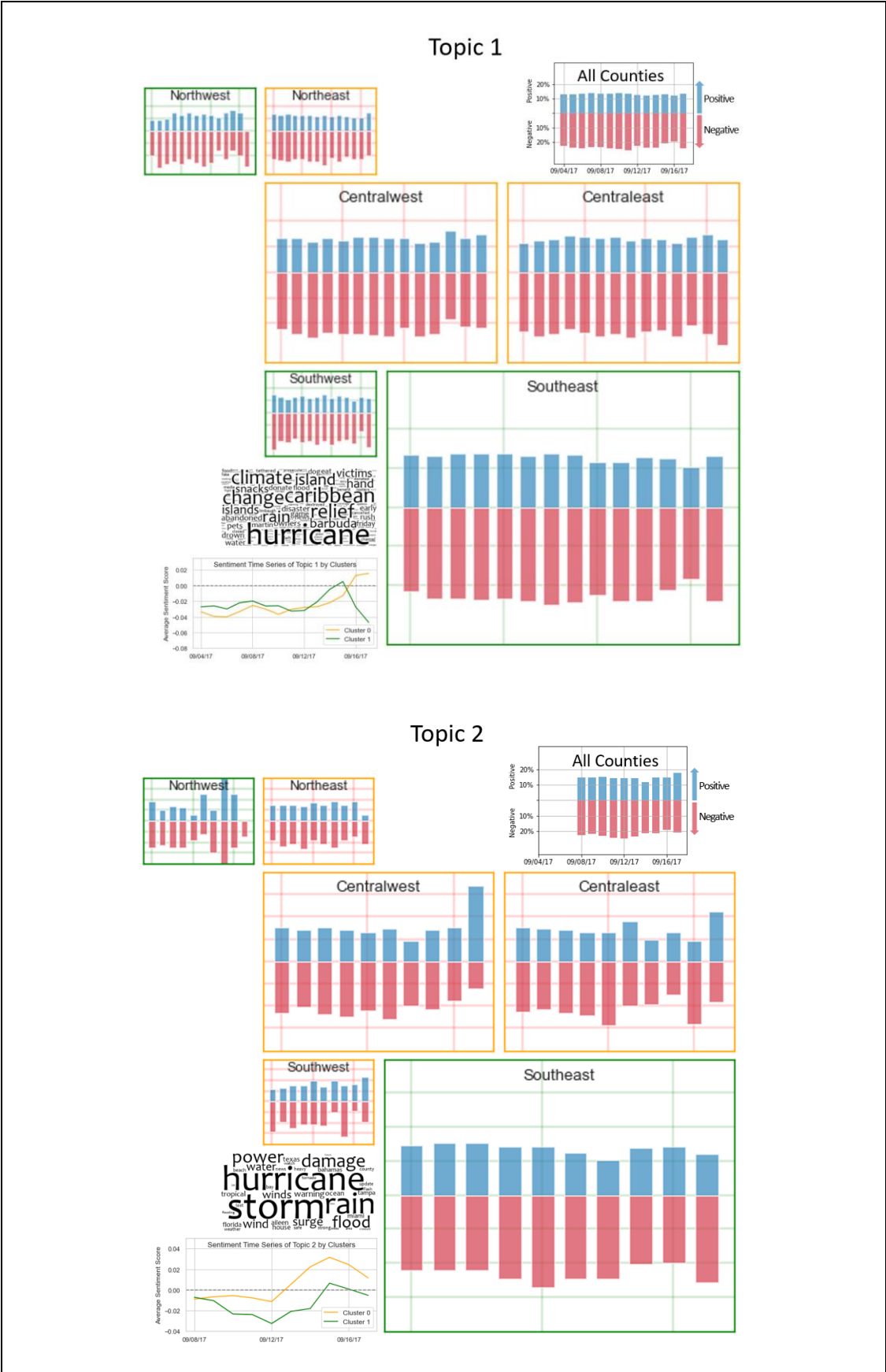


Figure A-5. Time series of sentiment distribution by regions for Topic 1 (upper) and Topic 2 (lower)