EXPRESSIVE FORMS OF TOPIC MODELING TO SUPPORT DIGITAL HUMANITIES

Samah H. Gad

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
Computer Science and Applications

Naren Ramakrishnan, Chair
Andrea L. Kavanaugh
Christopher L. North
Eli Tilevich
Niklas L. Elmqvist

September 8, 2014
Blacksburg, Virginia

Keywords: Topic Modeling, LDA, Time Series Segmentation, Visual Analytics

Copyright ©2014, Samah H. Gad
Unstructured textual data is rapidly growing and practitioners from diverse disciplines are experiencing a need to structure this massive amount of data. Topic modeling is one of the most used techniques for analyzing and understanding the latent structure of large text collections. Probabilistic graphical models are the main building block behind topic modeling and they are used to express assumptions about the latent structure of complex data. This dissertation addresses four problems related to drawing structure from high-dimensional data and improving the text mining process.

Studying the ebb and flow of ideas during critical events, e.g. an epidemic, is very important to understanding the reporting or coverage around the event or the impact of the event on the society. This can be accomplished by capturing the dynamic evolution of topics underlying a text corpora. We propose an approach to this problem by identifying segment boundaries that detect significant shifts of topic coverage. In order to identify segment boundaries, we embed a temporal segmentation algorithm around a topic modeling algorithm to capture such significant shifts of coverage. A key advantage of our approach is that it integrates with existing topic modeling algorithms in a transparent manner; thus, more sophisticated algorithms can be readily plugged in as research in topic modeling evolves. We apply this algorithm to studying data from the iNeighbors system, and apply our algorithm to six neighborhoods (three economically advantaged and three economically disadvantaged) to evaluate differences in conversations for statistical significance. Our findings suggest that social technologies may afford opportunities for democratic engagement in contexts that are otherwise less likely to support opportunities for deliberation and participatory democracy. We also examine the progression in coverage of historical newspapers about the 1918 influenza epidemic by applying our algorithm on the Washington Times archives. The algorithm is successful in identifying important qualitative features of news coverage of the pandemic.

Visually convincing results of data mining algorithms and models is crucial to analyzing and
driving conclusions from the algorithms. We develop ThemeDelta, a visual analytics system for extracting and visualizing temporal trends, clustering, and reorganization in time-indexed textual datasets. ThemeDelta is supported by a dynamic temporal segmentation algorithm that integrates with topic modeling algorithms to identify change points where significant shifts in topics occur. This algorithm detects not only the clustering and associations of keywords in a time period, but also their convergence into topics (groups of keywords) that may later diverge into new groups. The visual representation of ThemeDelta uses sinuous, variable-width lines to show this evolution on a timeline, utilizing color for categories, and line width for keyword strength. We demonstrate how interaction with ThemeDelta helps capture the rise and fall of topics by analyzing archives of historical newspapers, of U.S. presidential campaign speeches, and of social messages collected through iNeighbors. ThemeDelta is evaluated using a qualitative expert user study involving three researchers from rhetoric and history using the historical newspapers corpus.

Time and location are key parameters in any event; neglecting them while discovering topics from a collection of documents results in missing valuable information. We propose a dynamic spatial topic model (DSTM), a true spatio-temporal model that enables disaggregating a corpus’s coverage into location-based reporting, and understanding how such coverage varies over time. DSTM naturally generalizes traditional spatial and temporal topic models so that many existing formalisms can be viewed as special cases of DSTM. We demonstrate a successful application of DSTM to multiple newspapers from the Chronicling America repository. We demonstrate how our approach helps uncover key differences in the coverage of the flu as it spread through the nation, and provide possible explanations for such differences.

Major events that can change the flow of people’s lives are important to predict, especially when we have powerful models and sufficient data available at our fingertips. The problem of embedding the DSTM in a predictive setting is the last part of this dissertation. To predict events and their locations across time, we present a predictive dynamic spatial topic model that can predict future topics and their locations from unseen documents. We showed the applicability of our proposed approach by applying it on streaming tweets from Latin America. The prediction approach was successful in identify major events and their locations.
Contents

1 Introduction
   1.1 Motivation and Research Questions ........................................... 3
   1.2 Contributions ............................................................................... 6
   1.3 Outline of the Dissertation ........................................................... 8

2 Datasets
   2.1 iNeighbors .................................................................................. 9
   2.2 Chronicling America Historical Newspapers .................................... 12
   2.3 Presidential Campaigns Press Releases .......................................... 15
   2.4 Tweets from Latin America ............................................................ 17

3 Survey of Related Research
   3.1 Temporal Topic Modeling .............................................................. 20
   3.2 Topic Modeling other Extensions ................................................... 22
      3.2.1 Topic Modeling for Short Text ............................................... 22
      3.2.2 Syntactic Topic Modeling ....................................................... 23
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.3 Sentiment Analysis and Topic Modeling</td>
<td>24</td>
</tr>
<tr>
<td>3.2.4 Author Topic Model</td>
<td>24</td>
</tr>
<tr>
<td>3.2.5 Spatial Topic Modeling</td>
<td>25</td>
</tr>
<tr>
<td>3.3 Temporal Text Visualization</td>
<td>26</td>
</tr>
<tr>
<td>4 Dynamic Temporal Topic Modeling</td>
<td>29</td>
</tr>
<tr>
<td>4.1 Segmentation Algorithm</td>
<td>30</td>
</tr>
<tr>
<td>4.2 Algorithm Applications</td>
<td>34</td>
</tr>
<tr>
<td>4.2.1 Bridging the Divide in Democratic Engagement: Studying Conversation Patterns in Advantaged and Disadvantaged Communities</td>
<td>34</td>
</tr>
<tr>
<td>4.2.2 Digging into Historical Newspaper Archives using Dynamic Temporal Segmentations over Topic Models</td>
<td>46</td>
</tr>
<tr>
<td>4.3 Summary</td>
<td>55</td>
</tr>
<tr>
<td>5 New Visual Analytic Representations</td>
<td>56</td>
</tr>
<tr>
<td>5.1 ThemeDelta Overview</td>
<td>57</td>
</tr>
<tr>
<td>5.1.1 Data Format</td>
<td>58</td>
</tr>
<tr>
<td>5.1.2 Implementation</td>
<td>58</td>
</tr>
<tr>
<td>5.2 ThemeDelta: Visual Representation</td>
<td>59</td>
</tr>
<tr>
<td>5.2.1 Visual Design</td>
<td>59</td>
</tr>
<tr>
<td>5.2.2 Interaction</td>
<td>61</td>
</tr>
<tr>
<td>5.2.3 Layout</td>
<td>62</td>
</tr>
<tr>
<td>5.3 Domain Specific Applications</td>
<td>64</td>
</tr>
</tbody>
</table>
8.1 Dynamic Temporal Segmentations over Topic Models . . . . . . . . . . . . . . . . 108
8.2 New Visual Analytics Representations . . . . . . . . . . . . . . . . . . . . . . . . 109
8.3 Dynamic Spatial Topic Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110
8.4 Predictive Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111

Bibliography 112
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Sample pages from <em>Bisbee Daily Review</em>-AZ (December 17, 1918), <em>the New York Tribune</em>-NY (September 01, 1918), and <em>Red Lake News</em>-MN (November 01, 1918).</td>
<td>2</td>
</tr>
<tr>
<td>2.1</td>
<td>i-Neighbors: Social networking service connecting residents of geographic neighborhoods [iNe, 2012].</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>Distribution of messages across neighborhoods.</td>
<td>12</td>
</tr>
<tr>
<td>2.3</td>
<td>Distribution of messages across neighborhoods.</td>
<td>13</td>
</tr>
<tr>
<td>2.4</td>
<td>Distribution of influenza reporting over the year of 1918 and 1919.</td>
<td>15</td>
</tr>
<tr>
<td>2.5</td>
<td>Reporting concentration across locations and time.</td>
<td>15</td>
</tr>
<tr>
<td>4.1</td>
<td>Contingency table used to evaluate independence of topic distributions for two adjacent windows [Gad et al., 2012].</td>
<td>30</td>
</tr>
<tr>
<td>4.2</td>
<td>Partial segmentation output from a low-poverty neighborhood.</td>
<td>40</td>
</tr>
<tr>
<td>4.3</td>
<td>Partial segmentation output from a high-poverty neighborhood.</td>
<td>41</td>
</tr>
<tr>
<td>4.4</td>
<td>Partial segmentation output from a low-poverty neighborhood.</td>
<td>41</td>
</tr>
<tr>
<td>4.5</td>
<td>Durations of segments in advantaged (low poverty) and disadvantaged (high poverty) neighborhoods.</td>
<td>42</td>
</tr>
<tr>
<td>4.6</td>
<td>Example clusters of discovered segments across neighborhoods.</td>
<td>43</td>
</tr>
</tbody>
</table>
4.7 Segmentation results for The Washington Times Influenza paragraphs from September 1918 to December 1918. ...................................................... 52

4.8 Segmentation results for The Washington Times front pages from September 1918 to December 1918. ...................................................... 52

5.1 ThemeDelta visualization for Barack Obama campaign speeches during the U.S. 2012 presidential election (until September 10, 2012). Green lines are shared terms between Obama and Romney. Data from the “The American Presidency Project” at UCSB (http://www.presidency.ucsb.edu/). ........................................ 57

5.2 Basic visual representation used by ThemeDelta. .............................................. 59

5.3 ThemeDelta visualization after performing a filtering operation, based on the keyword “energy”, in the visualization presented in 5.1. ................................. 62

5.4 Comparison of different stages of the layout sorting algorithm used for the ThemeDelta technique. ................................................................. 63

5.5 ThemeDelta visualization for Mitt Romney campaign speeches for the U.S. 2012 presidential election (as of September 10, 2012). Green lines are shared terms between Obama and Romney speeches. Data from the American Presidency Project at UCSB (http://www.presidency.ucsb.edu/). ........................................ 64

5.6 Result of searching for the word “watch” in low-poverty neighborhood. .......... 67

5.7 Partial output from a high-poverty neighborhood. .............................................. 68

5.8 Partial output from a low-poverty neighborhood. .............................................. 69

5.9 ThemeDelta visualization for newspaper paragraphs during the period September to December in 1918. Color transparency for different trendlines signify the global frequency for that keyword. ......................................................... 70

6.1 Graphical model representation of the DSTM for three consecutive time slices. .. 77
6.2 Perplexity as a function of number of topics. ................................. 84
6.3 Perplexity as a function of vocabulary size. ................................. 84
6.4 New York Tribune, NY DSTM Output. ................................. 90
6.5 The Evening Missourian, MO DSTM Output. ................................. 91
6.6 Bisbee Daily Review, AZ DSTM Output. ................................. 91
6.7 Tones distribution over Influenza reporting. ................................. 93
6.8 East coast newspapers discovered topics and locations grouped by tones. ................................. 94

7.1 Experimental setup for predicting topics and their locations from streaming data. . 100
7.2 Predicted topics and their locations from the 8th day of June 2013 . . . . . . . 105
7.3 Predicted topics and their locations from June 29th, 2013. . . . . . . . . . . . . . . . 105
List of Tables

2.1 The six neighborhoods studied in our experiments. . . . . . . . . . . . . . . . . . .12
2.2 Daily newspapers summary. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .16

4.1 Event Timeline created from Front Pages of The Washington Times (1918). . . . . . . .51

6.1 DSTM notation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .77

7.1 Sample topic-document(tweet) assignment . . . . . . . . . . . . . . . . . . . . . . . .103
7.2 Predicated topic assignment counts for June 8th, 2013. . . . . . . . . . . . . . . . .104
7.3 Predicated topic assignment counts for June 29th, 2013. . . . . . . . . . . . . . . . .106
Chapter 1

Introduction

Historians and humanists are rapidly embracing the notion of ‘big data’ [Grossman, 2012] as a context to pose and investigate their research questions. The application of algorithmic techniques enables them to systematically explore a broad repository of data and identify qualitative features of a phenomenon (response, sentiment, and associations) in the small scale as well as the genealogy of information flow in the large scale.

The field of humanities has traditionally relied on close reading of documents in a topic of interest. The increasing availability of electronic document archives and their rapid growth has ushered in the era of digital humanities, and what is referred to as distant reading. Distant reading entails the comprehension of literature ‘not by studying particular texts, but by aggregating and analyzing massive amounts of data [Schulz, 2011].’

A key area that can benefit from distant reading of hundreds of text is in the comprehension of newspaper coverage of significant events, such as the 1918 ‘Spanish’ flu. Understanding the coverage of reported infected locations across local and national newspapers (see Fig. 1.1) is a key step that can help us understand how news propagated through time and space in those early times, when newspapers were the only widely used information resource.

A different medium, also relevant to modern digital humanities research, spans personal
blogs, Facebook posts, tweets, product reviews, and any shared information online by organizations or individuals. Mining social media in any of its forms it very important for social science researchers for many different reasons. Text mining is the concept of deriving high-quality features from text [Hotho et al., 2005]. One of the currently most promising lines of research in text mining is topic modeling by the formalism of Latent Dirichlet Allocation (LDA) [Blei et al., 2003], where documents are modeled as distributions (mixtures) over topics, and topics in turn are distributions over the vocabulary used in the corpus. LDA is considered a generalization of Probabilistic Latent Semantic Analysis (PLSA), proposed by Hofmann [Hofmann, 1999a]. (The difference between LDA and PLSA is that the topics distributions in LDA are assumed to be distributed according to a Dirichlet prior.)

Through text mining, a great number of social theories can be examined. For example, the detection of deliberation and common interests can be compared across different groups with specific demographics. Blogs, Facebook feeds, and tweets are great venues for characterizing public interest and opinions about a specific issue.

In the rest of this chapter, the motivation behind each part of this dissertation along with the
specific research questions will be presented. Then, contributions of different parts will be clearly stated. The last section of this chapter is an outline for the rest of this dissertation.

1.1 Motivation and Research Questions

Classic topic modeling has been applied in a great number of fields. Extensions and modifications have also been proposed in the literature. Some added a temporal aspect to topic models and others added structure to the discovered topics. The previously mentioned applications were a great motivation to build on and extend the classic topic models. In this section the motivation behind each part of this dissertation will be discussed and a short overview will be provided. This dissertation is divided into four major parts: Dynamic Temporal Segmentations over Topic Models, new visual analytic representations, Dynamic Spatial Topic Models (DSTM), and predictive analysis.

Dynamic Temporal Segmentations over Topic Models: The first part, Dynamic Temporal Segmentations over Topic Models, is motivated by significant ongoing research in capturing the dynamic evolution of topics underlying a text corpora. Most of these efforts are focused on extending the classical probabilistic model of Latent Dirichlet Allocation (LDA) [Blei et al., 2003] to a time-indexed context. Our temporal topic modeling approach is differentiated by its emphasis on automatically identifying segments where topic distribution is uniform and segment boundaries around which significant changes are occurring. We embed a temporal segmentation algorithm around a topic modeling algorithm to capture such significant shifts of coverage. A key advantage of our approach is that it integrates with existing topic modeling algorithms in a transparent manner; thus, more sophisticated algorithms can be readily plugged in as research in topic modeling evolves.

New Visual Analytic Representations: Several visual analytic applications require the analysis of dynamically changing trends over time. Example contexts include studies of idea diffusion in scientific communities, the ebb and flow of news on global, national, and local levels, and the meandering patterns of communication in social networks. Trends, each representing a particular keyword or concept, that converge into topics at different points in time, then just as unpredictably
diverge into new defined topics at a later time, are key patterns of interest to an analyst. Both experts and casual users alike need mechanisms for understanding such evolving trends for analysis, prediction, and decision making.

We present THEMEDELTA, a visual analytics system for accurately extracting and portraying how individual trends gather with other trends to form ad hoc groups of trends at specific points in time. Such gathering is inevitably followed by scattering, where trends diverge or fork to form new groupings. Understanding the interplay between these two behaviors provides significant insight into the temporal evolution of a dataset. Existing visualization techniques such as ThemeRiver [Havre et al., 2002] and streamgraphs [Byron and Wattenberg, 2008] are aimed at capturing overall trends in textual corpora but fail to capture their branching and merging nature. Our ThemeDelta temporal topic modeling approach is differentiated by its emphasis on automatically identifying segments where topic distribution is uniform and segment boundaries around which significant changes are occurring.

**Dynamic Spatial Topic Models (DSTM):** Temporal topic models have become quite standardized [Blei and Lafferty, 2006, Wang and McCallum, 2006, AlSumait et al., 2008, Gohr et al., 2009, Zhang et al., 2010, Hoffman et al., 2010, Hong et al., 2011]. Spatial topic models capture the notion of location but thus far have used location as a proxy for similarity [Pan and Mitra, 2011, Wang et al., 2009] (i.e., words closer in space are more similar to each other). In modeling newspapers that report events from across the country, we require topic models to be decomposable into specific topics for specific locations which are then aggregated in different ways to form news stories. Modeling such decompositions and tracking their evolution over time leads to a holistic understanding of coverage of large-scale events such as the Spanish flu.

In the third part of this work, we propose a new dynamic spatial topic model (DSTM) that incorporate reporting locations of inferred topics, and captures their evolutions over time. Topics (distributions over terms) are associated with locations and documents are comprised of multiple topics, i.e., coverage of several locations. The main goal behind building this model is to assist in the comprehension of newspaper coverage of significant events, such as the 1918 ‘Spanish’ flu.
Understanding the coverage of reported infected locations across local and national newspapers is a key step that can help us understand how news propagated through time and space in those early times, when newspapers were the only widely used information resource.

**Predictive Analysis:** The fourth and last part of this work is concerned with enabling powerful models to predict future topics. Enabling DSTM for predictive analysis will allow us to predict what, where, and when a major event will happen. We adapted the work of [Wang et al., 2012] where the idea is to train a basic topic model (LDA) on past data, and to calculate a topic distribution transition parameter from discovered topics. This transition parameter is then used to predict future topic distributions for unseen data. The transition parameter needs to be updated every time new data is streamed. Limitations of this work stem from its reliance on the vanilla LDA formulation, i.e., a non-dynamic and non-spatial topic model. Second, updating the transition parameter is computationally intensive. In this part of the dissertation we overcome those drawbacks by training the model using our DSTM approach. The inherent dynamicism in our model circumvents the need to update the transition parameters explicitly. Furthermore, the use of DSTM over LDA enables predicting the locations of topics in addition to topics. We demonstrate the use of this approach in forecasting civil unrest events (including their locations) in Latin America.

In summary, the research questions that will be explored in the four different parts of this dissertation are:

1. **Dynamic Temporal Segmentations over Topic Models:**
   - How do we identify segment boundaries that detect significant shifts of topic coverage?

2. **New Visual Analytic Representations:**
   - How can a visual analytics tool based on the segmentation algorithm facilitate dataset exploration?

3. **Dynamic Spatial Topic Model:**
• How can we generalize the basic topic modeling framework to accommodate location and temporal distinctions in large document sets?

4. Predictive Analysis:

• How can we use the DSTM for predicting attributes of future events?

5. General research question:

• Will the above modifications and extensions to classic LDA-based topic modeling help extract greater information from data and improve the utility of the text mining process?

Our goal is to increase the expressiveness of topic models as a text analysis tool. Classic topic modeling only focuses on word/token level analysis. These modifications to LDA will embed more structure and render the discovered topics much meaningful. To support this claim the presented work will be applied on a number of applications.

1.2 Contributions

As presented earlier, this dissertation is divided into four major parts and each part has a set of contributions. The first part is the Dynamic Temporal Topic Modeling and our specific contributions in this part are:

• A time series segmentation algorithm where segment boundaries detect significant shifts of topic coverage. To this purpose, we embed a topic modeling algorithm inside a segmentation algorithm and optimize for segment boundaries that reflect significant shifts of topic content.

• A novel application to studying Internet use in communities using the i-Neighbors system. The voluntary participation of i-Neighbors users enables us to gain significant insight into questions of engagement and deliberation.
• Qualitative as well as quantitative summaries of distinctions observed between advantaged and disadvantaged communities. These results lead to an understanding of how engagement and deliberation practices relate to access and uses of new communication technologies.

• A novel application to understanding the progression in coverage about the 1918 influenza from historical newspapers and a successful application of our algorithm to archives of the Washington Times. By studying the ebb and flow of ideas in the Fall of 1918 we illustrate how our algorithm extracts important qualitative features of news coverage of the pandemic.

The second part relates to new visual analytics representations and our key contributions can be summarized as follows:

• We present a visual analytics system, ThemeDelta, for accurately extracting and portraying how individual trends gather with other trends to form ad hoc groups of trends at specific points in time. Such gathering is inevitably followed by scattering, where trends diverge or fork to form new groupings. Understanding the interplay between these two behaviors provides significant insight into the temporal evolution of a dataset.

• We demonstrated several potential usage scenarios for our novel ThemeDelta system. The scenarios are: historical U.S. newspaper data from four months in the year 1918 during the second wave of the Spanish flu pandemic; the similarities and differences in trends and themes being discussed by the two candidates in the U.S. 2012 presidential campaign; and social messages exchanged between virtual communities via the i-Neighbors web-based application [iNe, 2012]. These applications are intended to demonstrate that ThemeDelta provides an interesting insight into datasets not immediately apparent through other representations.

In the Dynamic Spatial Topic Model (DSTM), the third part of this thesis, our key contributions can be summarized as follows:

• DSTM is a true spatio-temporal model and enables disaggregating a newspaper’s coverage into location based reporting, and how such coverage varies over time.
• DSTM naturally generalizes traditional spatial and temporal topic models so that many existing formalisms are special cases of DSTM. Conceptually, DSTM is closest to author-topic models [Rosen-Zvi et al., 2004] but where the notion of author is instead replaced by location.

• We demonstrate a successful application of DSTM to multiple newspapers from the Chroncling America repository. We demonstrate how our approach helps uncover key differences in the coverage of the flu as it spread through the nation, and provide possible explanations for such differences.

The fourth and last part of this dissertation, Predictive Analytics, our main contribution is as follows:

• A predictive dynamic spatial topic model that can predict future topics and their locations from unseen documents by adapting the work proposed by [Wang et al., 2012] and overcoming two main drawbacks of their approach.

• We show the applicability of our proposed approach for unrest predication from Latin American tweets.

1.3 Outline of the Dissertation

The rest of this dissertation is organized as follows:

• Chapter 2: Datasets

• Chapter 4: Survey of Related Research

• Chapter 5: New Visual Analytic Representations

• Chapter 6: Dynamic Spatial Topic Model

• Chapter 7: Predictive Analysis

• Chapter 8: Conclusions
Chapter 2

Datasets

This chapter is dedicated to describing the different datasets used in the four parts of this dissertation. The work presented here will be applied on four different datasets. These datasets were collected from the following APIs: iNeighbors, Chronicling America, the US presidential campaign repository, and Twitter. In the Dynamic Temporal Segmentations over Topic Models part, the iNeighbors and Chronicling America datasets were used. To evaluate the applicability of the New proposed Visual Analytic Representation (ThemeDelta) we applied the system on the iNeighbors, Chronicling America, and presidential campaign datasets. In the Dynamic Spatial Topic Model part, the model was applied on partial datasets derived from Chronicling America dataset. For predictive Analysis approach evaluation, we used the Twitter dataset (comprising tweets from Latin America). In the following sections, we will review each dataset in details.

2.1 iNeighbors

The iNeighbors system, shown in Figure 2.1, was created as part of a university research project first run from the Massachusetts Institute of Technology and later from the University of Pennsylvania that has been operational since 2004 [Hampton, 2010]. The site allows anyone in the United States
or Canada to join and create a virtual community that matches their geographic neighborhood. Users who join the website agree to a Terms of Use, as approved by the Institutional Review Board (IRB). Through the Term of Use, users are informed that participation is voluntary and that logs of user activity would be recorded and analyzed. The iNeighbors project was designed as a naturalistic experiment; there was no attempt to provide training or to encourage any individual user or community to participate. The website offers the following services:

- Discussion forum / email list: each neighborhood has a discussion forum that allows users to contribute and comment by email.
- Directory: a list of all group members and their profile information.
- Events calendar: a group calendar.
- Photo gallery: a group photo gallery.
- Reviews: user contributed reviews of local companies and services.
- Polls: surveys administered to other group members.
• Documents: storage for shared documents and links.

As of 2012, the i-Neighbors website has attracted over 110,000 users who have registered over 15,000 neighborhoods. The size of each group and the number of active groups varies from month to month. In a typical month, over 1,000 neighborhoods are active and over 7,000 unique messages are collectively contributed to neighborhood discussion forums, which in turn are viewed over 1 million times. This analysis focuses on the adoption pattern of the most active i-Neighbors communities, based on measures of the concentration of poverty, and the content of messages contributed to their respective discussion forums.

The percentage of families below the poverty level in geographic areas represented by the 20 most active i-Neighbors groups, shown in Figure 2.2, ranges from a low of 3.2% to a high of 47.6%. 40% of the most active neighborhoods are in areas of concentrated poverty. Given that 15% of Americans live below the poverty level [Kneebone and Nadeau, 2011], that 40% of the most active i-Neighbors groups are in areas where more than 20% of families are in poverty indicates adoption by high poverty neighborhoods at a higher rate than would be expected at random.

In this dataset, we ranked neighborhoods based on the number of unique comments that members posted to their neighborhood’s discussion forum over a one year period that started on October 1, 2010. For each neighborhood group, we identified the poverty rate, as defined by the US Census [cen, 2012], based on Census tract data collected as part of the 2009 American Community Survey (US Census Bureau). In Figure 2.3, the same neighborhoods shown in Figure 2.2 were rearranged based on poverty level. While recognizing that the selection of any absolute threshold will have its shortcomings, consistent with previous research, we used a 20% poverty rate as an indicator of an area of high-poverty [Kneebone and Nadeau, 2011].

We limited the scope of this dataset to the three most active i-Neighbors groups above our 20% poverty level threshold, and the three most active below the threshold. While we recognize that there are a number of potential sampling approaches, including sampling groups from similar or diverse geographic areas, we chose to maximize the available data for topic modeling. However, our approach also served to provide a sample that was geographically diverse, with the six groups
Table 2.1: The six neighborhoods studied in our experiments.

<table>
<thead>
<tr>
<th>Neighborhood ID</th>
<th>Number of Members</th>
<th>Number of messages</th>
<th>State</th>
<th>Poverty</th>
</tr>
</thead>
<tbody>
<tr>
<td>High1</td>
<td>440</td>
<td>2122</td>
<td>Ohio</td>
<td>47.60%</td>
</tr>
<tr>
<td>High2</td>
<td>334</td>
<td>3466</td>
<td>New York</td>
<td>26.30%</td>
</tr>
<tr>
<td>High3</td>
<td>539</td>
<td>2969</td>
<td>Maryland</td>
<td>24.90%</td>
</tr>
<tr>
<td>Low3</td>
<td>378</td>
<td>2472</td>
<td>Texas</td>
<td>6.60%</td>
</tr>
<tr>
<td>Low2</td>
<td>324</td>
<td>3534</td>
<td>Georgia</td>
<td>3.90%</td>
</tr>
<tr>
<td>Low1</td>
<td>371</td>
<td>2523</td>
<td>North Carolina</td>
<td>3.20%</td>
</tr>
</tbody>
</table>

Figure 2.2: Distribution of messages across neighborhoods.

used for our topic analysis representing six different U.S. States as shown in Table 2.1.

2.2 Chronicling America Historical Newspapers

Chronicling America, which is sponsored by the National Endowment for the Humanities and the Library of Congress, is a great example of an open source digital library of historical newspapers. It provides an Internet-based, searchable database of historical U.S. newspapers. The website is maintained by the National Digital Newspaper Program (NDNP). Example newspapers included in
this dataset are: The Washington Times (Washington, DC), Evening Public Ledger (Philadelphia, PA), The Evening Missourian (Columbia, MO), El Paso Herald (El Paso, TX), and The Holt County Sentinel (Oregon, MO). Data collected from these newspapers are stored as pages. Each page has a record in the dataset and the following information is available for each page: page OCR text, page number, newspaper name, and publish date.

We built this dataset by crawling the publicly accessible archive of Chronicling America Website. During the period we are interested in there were 104 newspapers available. The focus of this work is on the 1918 influenza epidemic. For this purpose we extract all paragraphs that contain one or more of the following words: influenza, grip, flu, epidemic, and grippe. Several sub-datasets were extracted from this dataset.

One sub-dataset was focused on 14 daily newspapers and extracted the influenza related paragraphs from them. Those paragraphs were extracted by searching the OCR text of newspapers pages. A summary for the daily newspapers we included in this sub-dataset is shown in Table 2.2. In the table, the period column indicates the duration in which Chronicling America provide data for a specific newspaper. The pages column provide the number of pages available for a newspaper. The number of pages that do contain one or more of the filtering keywords is shin in the relative pages
column. The last column is paragraphs, and it is a summary for the number of paragraphs extracted from a specific newspaper during the January 1918 to December 1919 period. The paragraph’s extraction from the daily newspapers resulted in 47,650 paragraphs.

For another sub-dataset, we ran a location detection script to label paragraphs with locations mentioned within their text. We provide the script with a list of all the cities and counties in all 50 USA states and military camps. We discarded paragraphs without location mentions. Paragraphs here are considered documents, and they are composed of five sentences. The five sentences are the result of including two sentences before and two sentences after the main sentence that contain one or more of the filtering keywords. Stop words, punctuation, and non-alphabetic characters were removed from paragraphs. Then we divide the dataset of extracted paragraphs into months. As a result, our data consist of 24 time slices from two years worth of data. Time slice sizes should vary based on the application. In a historical newspaper dataset, monthly time slices are appropriate because the news did not travel as fast as today’s news and because we are interested in major events that do have a monthly granularity. The resulting datasets will act as a stand alone dataset, one for each month.

Figure 2.4 shows the distribution of influenza reporting in the west, midwest, and east sides of the county over the year 1918 and 1919. Figure 2.5 displays the concentration of reporting for three different parts of the country. Columns in this grid represent the reporting percentage with respect to the other parts of the country. For each part, we used three different shades of the same color to show different levels of concentration. Concentration levels ranged from low to high and represented by light to dark shades of the same color. From this grid, we concluded that the midwest has a stable reporting on influenza compared to the east and west. The concentration of reporting in east, midwest, and west around the peaks of influenza confirms with the influenza spread. During the September 1918 and October 1918 the east was reporting more than the midwest and the west. The midwest reporting started to rise in November after a low concentration through previous months. Similarly, the west reported with a high concentration in November 1918 and continued with the same concentration through January 1919.
2.3 Presidential Campaigns Press Releases

Political speeches, especially during an election campaign, are particularly interesting document collections to analyze because the political discourse tends to change and evolve as different candidates respond and challenge each other over the course of the campaign.

The U.S. presidential election takes place every four years (starting in 1792) in November (the 2012 election day is November 6), and is an indirect vote on members of the U.S. Electoral College, who then directly elect the president and vice president. Electoral College members can vote for any eligible candidate, but are typically pledged to a particular candidate that has been officially nominated by a political party.
Table 2.2: Daily newspapers summary.

<table>
<thead>
<tr>
<th>East Coast Newspapers</th>
<th>Period</th>
<th>Pages</th>
<th>Relative Pages</th>
<th>Paragraphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>New-York tribune. (New York, NY)</td>
<td>1842-1922</td>
<td>313953</td>
<td>2768</td>
<td>1586</td>
</tr>
<tr>
<td>The Washington times. (Washington, DC)</td>
<td>1894-1939</td>
<td>143520</td>
<td>2755</td>
<td>9764</td>
</tr>
<tr>
<td>Evening public ledger. (Philadelphia, PA)</td>
<td>1914-1942</td>
<td>57602</td>
<td>2693</td>
<td>2230</td>
</tr>
<tr>
<td>The sun. (New York, NY)</td>
<td>1859-1920</td>
<td>225723</td>
<td>3027</td>
<td>5794</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Midwest Newspapers</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The Evening Missourian. (Columbia, MO)</td>
<td>1917-1920</td>
<td>4294</td>
<td>661</td>
<td>2518</td>
</tr>
<tr>
<td>El Paso herald. (El Paso, TX)</td>
<td>1901-1931</td>
<td>55154</td>
<td>2502</td>
<td>5096</td>
</tr>
<tr>
<td>The Corpus Christi caller. (Corpus Christi, TX)</td>
<td>1918-1987</td>
<td>8521</td>
<td>1087</td>
<td>2562</td>
</tr>
<tr>
<td>The Bismarck tribune. (Bismarck, ND)</td>
<td>1873-Current</td>
<td>15437</td>
<td>1044</td>
<td>2844</td>
</tr>
<tr>
<td>Tulsa daily world. (Tulsa, Indian Territory, OK)</td>
<td>1905-1919</td>
<td>38379</td>
<td>2020</td>
<td>5116</td>
</tr>
<tr>
<td>The Bemidji daily pioneer. (Bemidji, MN)</td>
<td>1904-1971</td>
<td>30037</td>
<td>891</td>
<td>2356</td>
</tr>
<tr>
<td>The evening herald. (Albuquerque, NM)</td>
<td>1914-1922</td>
<td>23714</td>
<td>1719</td>
<td>3600</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>West Coast Newspapers</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rogue River courier. (Grants Pass, OR)</td>
<td>1913-1918</td>
<td>9093</td>
<td>224</td>
<td>494</td>
</tr>
<tr>
<td>The Tacoma times. (Tacoma, WA)</td>
<td>1903-1949</td>
<td>25172</td>
<td>187</td>
<td>336</td>
</tr>
<tr>
<td>Bisbee daily review. (Bisbee, AZ)</td>
<td>1901-1971</td>
<td>52713</td>
<td>998</td>
<td>3354</td>
</tr>
</tbody>
</table>

In 2012, the Republican and Democratic (the two dominant parties in the United States, representing conservative vs. liberal agendas) conventions were held on the weeks of August 27 and September 3, respectively. The two opposing candidates were Republican nominee Mitt Romney, and Democratic nominee Barack Obama (incumbent President of the United States).

In collecting data for the United States presidential election, campaign speech transcripts for both candidates from the UCSB American Presidency Project were collected. For Mitt Romney, transcripts from 46 speeches over a 62-week period were used: from announcing candidacy on June 2, 2011, to September 10, 2012. This corpus included speeches from both the Republican primary election (settled on May 14, 2012 as the main competing nominee Ron Paul withdrew). For Barack Obama, transcripts from 41 speeches over a 7-week period were used: July 5, 2012 to September 10, 2012. This time period was much shorter due to the UCSB website only containing campaign speeches from July and onwards (earlier speeches were presumably given in an official capacity as sitting president). Significantly, no speeches were included from the Democratic primary elections, which Barack Obama secured on April 3, 2012.
2.4 Tweets from Latin America

This dataset consists of tweets collected from Latin American countries. Examples of countries included in this dataset are Brazil, Honduras, Colombia, Mexico, El Salvador, Costa Rica, Guatemala, Chile, Paraguay, Argentina, Venezuela, and Ecuador. This data was provided by EMBERS (Early Model-Based Event Recognition with Surrogates), a Virginia Tech project funded by the Intelligence Advanced Research Projects Activity (IARPA) OSI (Open Source Indicators) program. The main goal of EMBERS is to develop early warning indicators of significant population-level events such as civil unrest in countries of interest.

Tweets from these countries are filtered using a keyword list that help indicate the relationship of the tweet to any civil unrest event. Metadata about each tweet is captured. Metadata includes the tweet location, language, date, tweet author, and the number of author friends. Tweets are crawled daily and saved in a JavaScript Object Notation (JSON) format. Each tweet is represented as a JSON Object.

Tweets in this dataset were enriched with extra metadata. Each tweet JSON object include but not limited to the following tags:

- City: the city where the tweet originated from.
- Country: the country where the tweet originated from.
- Longitude and latitude: the geolocation of the tweet (if available).
- Interaction: this tag keeps track of the mentions in the tweet.
- Basis Enrichment: this tag includes automatic detected/extracted noun phrases and part of speech tags for each word in the tweet.
- Date: the date of the tweet.
- Twitter: this is the original tweet data that was crawled from Twitter without enrichment.
Along with the daily tweets, the EMBERS project also supplies a dataset of daily reported civil unrest events and their locations. The events serve as a ground truth source for validating the results of the developed models and algorithms. Reference events metadata include but not limited to the following information:

- Event Id: a unique id assigned for each event.
- Entry/Revision Date: when the event were recorded.
- Country: the country in-which this event took place.
- State: the state in-which this event took place.
- City: the city in-which this event took place.
- Population: type of population that is involved in this event. Population types include: General, Business, Labor, Agriculture, Religious, Medical, Education, and Legal.
- Date: the actual date of the event.
- Earliest reported date: the date the event were reported.
- News source: the source where the event were publicized.
- Headline: the headline used when the event was publicized.
- Event Description: a full description of the event.
Chapter 3

Survey of Related Research

This chapter serves as a survey for the related research to the different parts of this dissertation. The most used Topic Modeling approach was LDA, first presented by [Blei et al., 2003], where documents are modeled as distributions (mixtures) over topics, and topics in turn are distributions over the vocabulary used in the corpus.

LDA assumes that documents are generated in two stages: randomly specify a topic-terms distribution. Then, for each word in a document: randomly choose a topic from the chosen distribution and then randomly choose a word from the topic distribution. The generative process for LDA can be expressed as a joint probability distribution over the observed documents. Words in documents are the only observed valuables and the topic-term distribution, document-topic distribution, topic assignment, topics priors, and document prior are all latent variables in the LDA model.

The main goal of this model is to calculate the posterior probability of topic assignment for each word in the observed document conditioned on all other variables. This probability is very hard to calculate. To solve the posterior probability they use Collapsed Gibbs Sampling.

There are three main research areas surveyed here: Temporal Topic Modeling, Topic Modeling Extensions, and Temporal Text Visualization. In the following sections, each area will be
address in details.

## 3.1 Temporal Topic Modeling

Temporal topic modeling algorithms started to appear around 2006, most of them being generalizations of static topic models. The difference between our goals and those of previous work is that we aim to automatically identify segment boundaries that denote shifts of coverage, and, in this manner, extract temporal relationships for examination. Therefore, we are not proposing a new topic model but instead proposing how we can “wrap” around existing topic modeling algorithms to segment time-stamped data.

Classical work in this space was done by Blei and Lafferty [Blei and Lafferty, 2006], who extended traditional state space models to identify a statistical model of topic evolution. They also developed techniques for approximating the posterior inference for detecting topic evolution in a document collection.

Documents in Blei and Lafferty model are modeled independently at each time slice. Modeled topics from time $t$ are evolved from topics modeled from $t-1$. They replaced the Dirichlet distribution used in classic LDA to draw the topic distribution over the vocabulary with a Gaussian distribution. The word in a document is the only observed variable in their model. Latent variables in their model are: topic assignment, topic-term distribution, document-topic distribution, and the Dirichlet prior for the document-topic distribution.

The generative process for documents at time $t$ is similar to the LDA generative process, but the difference is that here the priors for each topic are chained in a state space model and it evolve with Gaussian noise. Also, a document-topic distribution are drawn from a logistic normal distribution to express uncertainty over proportions. They presented a variational method for approximate posterior inference.

Wang and MacCallum [Wang and McCallum, 2006] proposed a non-Markov model for
detecting topic evolution over time. They assume that topics are associated with a continuous distribution over timestamps and that the mixture distribution over topics that represent documents is influenced by both word co-occurrence relationships and the document timestamp. In their model, thus, topics generate both observed timestamps and words. Iwata and Yamada [Iwata et al., 2010] also proposed a topic model that enables sequential analysis of the dynamics of multiple time scale topics. In their proposed model, topic distributions over words are assumed to be generated based on estimates of multiple timescale word distributions of the previous time period. Finally, Wang and Blei [Wang et al., 2008] have recently proposed a model that replaces the discrete state space that was originally proposed by Blei and Lafferty [Blei and Lafferty, 2006] with a Brownian motion law [Lawler, 1995] to model topic evolution. They assume that topics are divided into sequential groups so that topics in each slice are assumed to evolve from the previous slice. This line of research has been extended to mining text streams, e.g., as done in [Wang et al., 2013]. Here, the authors study the problem of mining evolving multi-branch topic trees inside a text stream by proposing an evolutionary multi-branch tree clustering method. In their method, they adopt Bayesian rose trees to build multi-branch trees and use conditionals prior over tree structures to keep the information from previous trees as well as maintain tree smoothness over time. To keep the consistency of trees over time, they define a constraint tree from triples and fans to compute the tree structure differences.

Some recent papers have targeted the goal of modeling multiple information sources along with capturing topic evolution. Zhang et al. [Zhang et al., 2010] have proposed an evolutionary hierarchal Dirichlet process (EvoHDP) model which extends the hierarchical Dirichlet process (HDP) to take time into account [Teh et al., 2004]. Inference of EvoHDPs is conducted through a cascade Gibbs sampling strategy. Hong, Dom, Gurumurthy, and Tsioutsiouliklis [Hong et al., 2011] have also addressed multiple streams and the temporal dynamics of topics detected from these streams. They tackle the multiple stream problem by allowing each text stream to have both local topics and shared topics. Each topic is associated with a function that characterize the topic popularity over time and this function is time-dependent.

Other work focused on multiple text streams. Wang et al. [Wang et al., 2009] and [Wang
et al., 2007] aim to “align” time series streams so as to identify correlated and/or common topics across disparate streams. Our algorithm was designed to analyze a single time-indexed corpus as opposed to multiple time (and asynchronous) text streams. The work of Leskovec et al. [Leskovec et al., 2009] strings together individual tweets (meme) into a thread. The authors place additional constraints in identifying these threads (e.g., that they originate in a single meme). The granularity of analysis in our work are clusters rather than tweets, and the desired output is rich scatter-gather relations between clusters rather than simple branching patterns. The work presented by Gao et al. [Gao et al., 2011] is the closest to our work. The authors conduct topic modeling, and investigate both scattering and gathering possibilities of cluster organization. The difference is that our work automatically determines segmentation boundaries where significant shifts of topic distributions occur whereas the work of Gao et al. incrementally clusters every time point separately and then aims to make splitting and merging decisions.

3.2 Topic Modeling other Extensions

3.2.1 Topic Modeling for Short Text

Short text (microblogs) like tweets and Facebook feeds are a great source for mining public opinions and interests. A number of research was done in order to merge the benefits of enabling the classic LDA to work with short text. An empirical study of topic modeling in microblogging environments was introduced by [Hong and Davison, 2010]. No changes to the topic models were introduced. They introduced ways to conduct existing topic models on short text. They demonstrated that aggregating short messages is a valid way to solve the document length dependency of topic models problem by conducting extensive qualitative and quantitative experiments on three proposed schemes based on classic LDA.

Another model, Temporal-LDA (TM-LDA), was proposed by [Wang et al., 2012] for mining streams of social text. The main idea was based on modeling the topics and topic transitions in data.
The proposed model detects transitions between topics by minimizing the prediction error on topic distribution in subsequent postings. It can also predict the expected topic distribution in future posts. A transition parameters updating algorithm was also developed to reach a more efficient prediction in online streaming settings.

### 3.2.2 Syntactic Topic Modeling

The first to take syntactic level information into account in topic modeling was [Boyd-Graber and Blei, 2008]. They developed a nonparametric Bayesian model of parsed documents. The topics discovered with their model were based on thematic and syntactic constraints. In their model words order was assumed to be generated based on their order in a parse tree.

Another research by Wallach et al. [Wallach, 2008] considered language and document structure in building their model. Low-level structures including word order and syntax were considered. Higher-level structures, such as relationships between documents were also considered. Latent topics were combined with information about document structure, ranging from local sentence structure to inter-document relationships. Three different structured topic models were introduced. The first is a topic-based language model that captures both word order and latent topics. The second is a dependency parsing topic model that is based on a Bayesian reinterpretation of a dependency parsing model. The third and last is a high-level relationships between documents topic model that uses a nonparametric Bayesian priors and Markov chain Monte Carlo methods to infer topic-based document clusters.

Guo et al. in [Guo and Ramakrishnan, 2010] also developed a Latent Dirichlet Allocation (LDA) that uses linguistic dependency information as replacement of the features extracted from Bag-of-Words (BOW) representations. They applied the proposed algorithm on movie reviews for spoiler detection.
3.2.3 Sentiment Analysis and Topic Modeling

Research on combining sentiment analysis and topic modeling can be divided into two main tracks. First, topic modeling was used to detect sentiment. Second, sentiment was extracted simultaneously while discovering topics.

Some research was doing both, for example, Mei et al. [Mei et al., 2007] proposed a probabilistic model called topic-sentiment mixture to capture topics and sentiments simultaneously. There is research that linked sentiment analysis with topic modeling, but from the perspective of using topic modeling for feature extraction to accurately detecting sentiments. Liu provides a survey on such work in his book chapter [Liu, 2010].

Opinion mining is a field that is highly related to sentiment analysis. Lu and Zhai, in [Lu and Zhai, 2008] were using expert reviews to mine text data for ordinary opinions. The main challenge they faced was the need to align ordinary opinions to an expert review and separate similar and supplementary opinions. For this propose, they used a semi-supervised topic modeling approach to solving these challenges.

Lin and He in [Lin and He, 2009] proposed joint sentiment/topic model (JST), a probabilistic modeling framework that is based on LDA. It can detect sentiment and topics simultaneously from text. They focused on document level sentiment an in order to detect sentiment they added sentiment detection layer on top of a classic LDA.

3.2.4 Author Topic Model

Rosen et al. [Rosen-Zvi et al., 2004] proposed a topic model that extract topics and authors information from a text collection. The model generates two major distributions: The first is the topic-terms distribution, which represents the topic distribution over the terms in the text collection vocabulary. The second is the author-topic distribution which is the author distribution over the discovered topics. In their model, the only observed variables are document authors and the words
in a document. According to their model, documents are generated as following: first, an author distribution over topics is randomly chosen. Then for each topic a distribution over the vocabulary distribution is randomly chosen. In a multi author document, the probability distribution over topics is a mixture of the distributions associated with the authors. Finally, for each word in a document: an author is randomly chosen, then the topic is randomly drawn from author-topic distribution. Then draw a word from the chosen topic. The main probability they are trying to calculate is the probability of assignment the word in a document to the topic and author respectively conditioned on all other variables. The conditional probability approximation in this work is done using a Markov chain Monte Carlo algorithm. They applied their model on three datasets: emails from a corporation, abstracts from CiteSeer library, and papers from Neural Information Processing Conference. The quantitatively evaluated their model by calculation perplexity and comparing their models to other models.

### 3.2.5 Spatial Topic Modeling

Space is another aspect that has caught the interest of topic modeling researchers. Two algorithms were introduced by Pan et al. [Pan and Mitra, 2011] for spatio-temporal event modeling from the text. The first was a three-Step LDA algorithm, based on combining the popular LDA model with temporal segmentation and spatial clustering. Here, LDA is first used to obtain document distributions over topics, and temporal and spatial references in documents within each topic are then resolved. The second algorithm, a space-time LDA algorithm was based on work introduced in [Wang et al., 2009]. This approach was originally used to encode spatial structure among visual words for image segmentation. The generative procedure of this algorithm was partitioning visual words that are close in space into the same documents. The main difference between this work and our DSTM is that our proposed DSTM uses a classical discrete time approach to capturing topic evolution, but words are reliant on both topic distributions and location distributions.
3.3 Temporal Text Visualization

Text visualization uses interactive visual representations of text that go beyond merely the words themselves to show important features of a large dataset [Don et al., 2007]. The simplest feature is word frequency; tag clouds (or word clouds) [Hearst and Rosner, 2008, Lohmann et al., 2009] are accordingly the most common text visualization technique, and use graphical variables like font, color, weight, and the size of the word to convey its importance in the source text. Variations include Wordle [Viégas et al., 2009] and ManiWordle [Koh et al., 2010] for producing more compact and aesthetically pleasing layouts.

More advanced text visualization techniques convey not only word frequency, but also structural features of the text corpus (or corpora), often using a graph structure. Examples include WordBridge [Kim et al., 2011], Phrase Nets [van Ham et al., 2009], and GreenArrow [Wong et al., 2005], which all use a variant of a dynamic graph to convey the structure in the document. Clustered word clouds [?, Hassan-Montero. and Herrero-Solana, 2006] take another approach by using word structure to modify the layout in a tag cloud, which is similar to the themescape representation used by IN-SPIRE [Wise et al., 1995].

Text data may often be analyzed over time to expose aspects of the data that evolved during a time period; for example, recent studies have highlighted the importance of time and causality in investigative analysis [Kwon et al., 2012]. While less saturated than general text visualization, there exists several pieces of work that visualize text over time in this field.

Perhaps one of the earliest and most seminal of these is ThemeRiver [Havre et al., 2002], which essentially can be thought of as a horizontal arrangement of a large number of one-dimensional tag clouds over time; the keywords become bands in a stacked graph evolving on a timeline. Byron and Wattenberg [Byron and Wattenberg, 2008] later formalized these representations into so-called streamgraphs, and applied it to many other types of data.

Several approaches exist that are similar to the basic ThemeRiver stacked chart. Tufte [Tufte, 1983] showed an illustration of changing themes of music through the ages. NewsLab [Ghoniem
et al., 2007] applies a ThemeRiver representation to visualize a collection of thousands of news videos over time. NameVoyager [Wattenberg, 2006] uses streamgraphs to show temporal frequencies of baby names. TIARA [Wei et al., 2010] blends tag clouds onto temporal stacked charts for important themes. However, because stacked charts group themes into a single visual entity, none of these techniques is capable of conveying the structural features of the corpus.

A select few techniques provide both temporal and structural information. Parallel tag clouds (PTCs) [Collins et al., 2009b] integrate one-dimensional tag cloud layouts on a set of parallel axes, each which could potentially be used for a time or date. However, while visually similar to our ThemeDelta technique, PTCs do not explicitly convey the grouping of words into topics. Another similar design is TimeNets [Kim et al., 2010], which uses colored lines to show grouping over time for genealogical data. In fact, “Movie Narrative Charts” (comic 657) of the web comic XKCD\(^1\) also uses sinuous lines to convey groupings of characters in time and space for famous movies. Finally, Turkay et al. [Turkay et al., 2011] present two techniques for visualizing structural changes of temporal clusters that are remniscient of ThemeDelta; while not specifically designed for text, the clusters used in their work could very well stem from textual corpora. ThemeDelta takes a similar visual metaphor, but focuses on text and is intrinsically tied to the scatter/gather temporal segmentation component as well.

A very relevant work is ParallelTopics, a visual analytics system that integrates LDA with interactive visualization [Dou et al., 2011]. The system uses the parallel coordinate metaphor to present document-topic distributions, with applications to exploring National Science Foundation awarded proposals, VAST publications, as well as tweets. This system presents the underlying probabilistic distributions in the LDA model from a temporal perspective using multiple aggregation strategies and interactions. The system can capture the topics and their evolution over time, but only using fixed time frames. In contrast, our ThemeDelta approach discovers time frames automatically based on topics reorganizations across time.

Finally, TextFlow [Cui et al., 2011] is perhaps the closest related work to ThemeDelta, and

\(^1\)http://www.xkcd.com/
uses tightly integrated visualization and topic mining algorithms to show an evolving text corpus over time. However, whereas we draw upon the same basic visual representation as TextFlow, our focus in this work is segmenting time based on topic shifts and then interfacing with standard topic modeling using a novel algorithm. Furthermore, ThemeDelta does not aggregate keywords into stacks or glyphs, and puts more emphasis on interactive layout.
Chapter 4

Dynamic Temporal Topic Modeling

The main goal behind the work presented here is to examine the ability to identify segment boundaries that detect significant shifts of topic coverage. The motivation that drove this work is the significant research done in detecting topic evolution in a text corpora. Research in the literature focused on extending the Latent Dirichlet Allocation (LDA), the classic topic model proposed by [Blei et al., 2003].

In this chapter, we will present a time-series segmentation algorithm that identify segmentation boundaries. These segmentation boundaries are detected when a significant shift of topics coverage occurs. To detect shifts in topics, we embed a topic modeling algorithm within a segmentation algorithm. To contrast our approach with the work mentioned in the literature 3.1, the goal is to not simply to track the temporal evolution of topics, but to identify segments that denote significant shifts in their content (distribution).

We use the algorithm to study Internet use in advantaged and disadvantaged communities. The dataset used for this application was the i-Neighbors dataset 2.1 We also applied the algorithm on paragraphs extracted from The Washington Times newspaper. The newspaper data was extracted from the Historical Newspaper dataset presented in 2.2. This application focused on studying the coverage of the influenza epidemic in 1918.
4.1 Segmentation Algorithm

Our segmentation algorithm expects the input data to be in a bag-of-word format. The preprocessing needed is thus to tokenize the text into individual words, followed by standard processing steps such as: lower case conversion, stemming, stop words removal, spell checker, and punctuation removal. The main task of the segmentation algorithm is to automatically partition the total time period defined by the documents in the collection such that segment boundaries indicate important periods of temporal evolution and re-organization.

![Segmentation Diagram]

Figure 4.1: Contingency table used to evaluate independence of topic distributions for two adjacent windows [Gad et al., 2012].

The algorithm moves across the data by time and evaluates two adjacent windows assuming a given segmentation granularity (e.g., discrete days, weeks, or months). This granularity varies from application to another and it is decided by domain experts. We evaluate adjacent windows by comparing their underlying topic distributions and quantifying common terms and their probabilities.

We chose to quantify common terms based on the overlap between them. The overlap can be captured using a contingency table. Figure 4.1 shows a simplified example of two segments, each comprising three topics and the corresponding contingency table measuring the overlap between these distributions. For example, topic 1 ($Z_1$) in segment 1 and topic 1 ($Z'_1$) in segment 2 overlap in $w_1$ and $w_6$. This resulted in adding the count 2 in the contingency table cell that corresponds to the overlapping cell between the two topics from the two segments. We would like the topic models of the two adjacent windows to be maximally independent, which will happen if the table entries
are near uniform.

Formally, given the input data to be indexed over a time series $T = \{t_1, t_2, \ldots, t_t\}$, the segmentation problem we are trying to tackle is to express $T$ as a sequence of segments or windows: $S_T = (S^a_{t_1}, S^b_{t_{n+1}}, \ldots, S^d_{t_k})$ where each of the windows $S^e_{t_s}, t_s \leq t_e$ denotes a contiguous sequence of time points with $t_s$ as the beginning time point and $t_e$ as the ending time point.

Each window $S^e_{t_s}$ has a set of topics that is discovered from the set of documents that fall within this window. The topics are discovered by applying LDA (Latent Dirichlet Allocation) [Blei et al., 2003]. Applying this algorithm will result in two main distributions: document-topic distribution (representing the distribution of the discovered topics over the documents) and topic-terms distribution (representing the distribution of the discovered topics over the vocabulary).

Topics within each window is represented as $S^e_{t_s} = \{z_1, z_2, \ldots, z_n\}$ where $n$ is the number of top topics $z$ discovered. Each topic is represented by a set of terms $w$ as follow: $z_i = \{w_1, w_2, \ldots, w_m\}$ where $m$ is the number the top terms extracted from the topic-terms distribution resulted from applying LDA on the documents within a window. Number of top topics $n$ and top terms representing a topic $m$ vary from application to another.

We represent two adjacent windows as $S^e_{t_{s_1}}$ and $S^e_{t_{s_2}}$. To evaluate two adjacent windows, we construct the contingency table for two windows. The contingency table is of size $r \times c$ where rows $r$ denote topics in one window and columns $c$ denote topics in the other window. Entry $n_{ij}$ in cell $(i, j)$ of the table represents the overlap of terms between topic $i$ of $S^e_{t_{s_1}}$ and topic $j$ of $S^e_{t_{s_2}}$.

We used a contingency table because it enable the replacement of LDA with any emerging topic modeling variants. As presented in [M. Shahriar Hossain, 2013] we can embed any vector quantization clustering algorithm in a contingency table framework. For instance, distributions inferred from a more sophisticated model can be compared using the contingency table formulation introduced here.

Then to check the uniformity of the table, three steps should be accomplished:
First, calculate the following two quantities:

- Column-wise sums \( n_i = \sum_j n_{ij} \)
- Row-wise sums \( n_j = \sum_i n_{ij} \)

These two quantities will be used to quantify the overlap between the topics discovered from two adjacent windows. In our implementation for this step, each topic is represented by its top assigned terms. The contingency table is created from these terms (here we chose 20 terms and the choice of the number of terms is inherently heuristic and specific to the application). A probabilistic similarity measure such as the KL- or JS-divergence between the distributions being compared is another possibility.

Second, we define two probability distributions, one for each row and one for each column:

\[
p(R_i = i) = \frac{n_{ij}}{n_i}, (1 \leq j \leq c) \tag{4.1}
\]

\[
p(C_j = j) = \frac{n_{ij}}{n_j}, (1 \leq i \leq r) \tag{4.2}
\]

Third, we calculate the objective function \( F \) to capture the deviation of these row-wise and column-wise distributions with regard to the uniform distribution.

The objective function is defined as follows:

\[
F = \frac{1}{r} \sum_{i=1}^r D_{KL}(R_i \| U(\frac{1}{c})) + \frac{1}{c} \sum_{j=1}^c D_{KL}(C_j \| U(\frac{1}{r})) \tag{4.3}
\]

where

\[
D_{KL}(P \| Q) = \sum_i P(i) \log \frac{Q(i)}{P(i)} \tag{4.4}
\]
This objective function can reach a local minimum, which is acceptable given that we are trying to segment time based on shifts in topics and this approach captures the first shift in topics (as opposed to detecting an optimal segmentation which would require a more exhaustive search through breakpoint layouts).

**Algorithm 1. Topic Modeling Based Segmentation**

**Input:** \( T = \{t_0, t_1, t_2, t_3, \ldots, t_t\} \)
- \( x = \) min. window size.
- \( y = \) max. window size.

**Output:** \( S_T = \{\} \) //Set of all segments between \( t_0 \) and \( t_t \)

\( W1Start = t_0 \)
\( W1Size = x \)

\( F = \) Initialize objective function with a large number.

**while** \( W1Start + W1Size + x \leq t_t \) and \( W1Size \leq y \) **do**

//x is added to W1 to take into account the data availability for W2.

- Conversion = False
- \( W2Start = W1Start + W1Size + 1 \text{day} \)
- \( W2Size = x \)

**while** \( W2Start + W2Size \leq t_t \) and \( W2Size \leq y \) **do**

- Apply LDA separately on W1 and W2
- Calculate \( F' \) for W1 and W2

  **if** \( F' > F \) **or** \( W1Size == y \) **or** \( W2Size == y \) **do**

    //Conversion or max. window size limit reach.
    - Add W1 and W2 to \( S_T \)
    - \( W1Start = W2Start + W2Size + 1 \text{day} \)
    - \( W1Size = x \)
    - Conversion = True

**Break**

- \( F = F' \)
- \( W2Size += x \) //Expand W2.

**if** !Conversion **do**

- \( W1Size += x \)

**if** leftover data exists **do**

  //leftover data starts at \( W1Start \) and ends at \( t_t \).
  - Apply LDA on leftover data.
  - Add window of leftover data to \( S_T \).

**return** \( S_T \)

---

Here, \( D_{KL} \) denotes the KL-divergence that is used to calculate the distance between the
row-wise and the uniform distribution. Likewise, it is used to calculate the distance between the column-wise distributions and the uniform distribution. Then the values resulting from using the $D_{KL}$ will be used in calculating the objective function $F$.

The algorithm repeats the above mentioned steps for all permutations of the two sliding window sizes. The goal is to minimize $F$, in which case the distributions observed in the contingency table are as close to a uniform distribution as possible, in turn implying that the topics are maximally dissimilar.

There are two stopping conditions for this algorithm: (1) if conversion of $F$ is achieved, or (2) the maximum size for both windows was achieved. Detailed description of the algorithm is shown in Algorithm 1. In the following section, two applications for this algorithm will be presented.

## 4.2 Algorithm Applications

### 4.2.1 Bridging the Divide in Democratic Engagement: Studying Conversation Patterns in Advantaged and Disadvantaged Communities

This work was done as a collaboration with Naren Ramakrishnan (Department of Computer Science, Virginia Tech), Keith N. Hampton (School of Communication and Information, Rutgers University) and Andrea Kavanaugh (Department of Computer Science, Virgina Tech). And was published in the ACM Social Informatics 2012 [Gad et al., 2012].

The Internet offers opportunities for informal deliberation, and civic and civil engagement. However, social inequalities have traditionally meant that some communities, where there is a concentration of poverty, are both less likely to exhibit these democratic behaviors and less likely to benefit from any additional boost as a result of technology use. We argue that some new technologies afford opportunities for communication that bridge this divide. Using temporal topic modeling,
we compare informal conversational activity that takes place online in communities of high and low poverty. Our analysis is based on data collected through iNeighbors, a community website that provides neighborhood discussion forums. We examine the adoption of iNeighbors by poverty level, and apply our algorithm to six neighborhoods (three economically advantaged and three economically disadvantaged) and evaluate differences in conversations for statistical significance. Our findings suggest that social technologies may afford opportunities for democratic engagement in contexts that are otherwise less likely to support opportunities for deliberation and participatory democracy.

Democratic engagement, at both the individual and community levels, is one of the strongest predictors of well-being [Helliwell and Putnam, 2004]. While political behaviors, such as voting, are among the most studied aspects of democratic engagement, they are only a small subset of the behaviors that contribute to a democracy. Participation in a democracy involves more than the occasional selection of representatives. Citizens and their communities benefit from individual and collective action to address issues of common concern through activities outside of elections and government [Carpini and Keeter, 1996]. Participatory democracy includes a range of civic behaviors, including membership in institutions that address public issues, such as a neighborhood watch [Putnam, 2000], as well as civil behaviors, such as helping a neighbor in an emergency [Klinenberg, 2002]. These behaviors are intertwined with casual conversations, that, although not overtly deliberative or political, are a part of the “incomplete” [Fishkin and Stone, 1995] forms of political deliberation that are key to shaping social identities, friendships, and trust [Walsh, 1992]. This combination of informal participation and casual, public deliberation provides for the social mixing that is important for opinion formation, awareness of common interests, social tolerance, and the ability to act on collective goals [Dewey, 1927]. Unfortunately, like so many forms of democratic engagement, civic and civil behaviors and informal opportunities for deliberation are unequally distributed.

Civic and civil behaviors, including opportunities for informal deliberation, are stratified by class [Uslaner and Brown, 2005]. Those of lower income are significantly less likely to exhibit attitudes and behaviors for democratic engagement [Carpini and Keeter, 1996]. In addition,
inequality is not equally distributed across the country, but concentrate in geographic areas of concentrated disadvantage; neighborhoods that are high in poverty, racial segregation, and social problems, such as crime [Sampson, 2011]. The concentration of inequalities is associated with structural instability that reduce the ability of residents to form the local social bonds necessary for collective action [Sampson, 2011]. As a result, those communities with the greatest need for informal discussion and participatory democracy are typically those where it is most absent.

Research on the role of new information and communication technologies (ICTs) and democratic engagement have generally found positive relationships between exposure to online political information and democratic behaviors [Shah et al., 2005, Boulianne, 2009]. Participation in online activities that support informal deliberation, such as social networking services, has also been found to contribute to political participation [Hampton et al., 2011]. However, there is almost no evidence that the use of ICTs overcomes existing socioeconomic inequalities associated with democratic engagement [Hargittai and Shaw, 2011]. Indeed, there may be a “Matthew effect” [Merton, 1968], such that those who are already the mostly likely to express democratic behaviors gain further as a result of new ICTs, while those who have little gain little as a result of ICT use.

We argue an alternative theory. We believe that new ICTs, specifically social media, offer new affordances for group interaction, informal deliberation and democratic engagement [Kavanaugh, 2013]. Unlike some other Internet technologies, social media afford contact in contexts where individuals have a shared affinity – through geography, political interests, or other interest – but previously lacked the means or ease of access for connectivity (in-person or online). We focus on how these affordances reduce the cost of communication for urban communities with concentrated inequalities.

This reduction in the cost of communication helps residents overcome established structural barriers to social tie formation, informal deliberation and participatory democracy. The result is a set of opportunities for democratic engagement among people and in areas previously constrained by structural barriers to collective action. When such social media that are designed to bring local people together are made available to people in urban neighborhoods with high socioeconomic
inequalities, we expect to find democratic engagement that is as high as what is typical of areas where such inequalities are less concentrated.

Specifically, our goal is to study the adoption of a tool for informal deliberation at the neighborhood level and to compare conversation patterns across advantaged and disadvantaged communities based on their level of concentrated poverty. Our aim is to characterize differences in informal deliberation, if any, between these advantaged and disadvantaged neighborhoods, as well as to detect common interests between them. This will provide insight into how neighborhoods with different poverty levels use ICTs for informal deliberation.

In order to be able to detect deliberation and common interests, we applied our temporal segmentation algorithm. The objective of applying the algorithm is to detect segments where there are significant concordances of topics, but such that segment boundaries identify significant shifts in topics.

Once a neighborhood discussion is characterized in this manner, we can: compare the time duration of topics in neighborhoods with different poverty levels, identify differences in topics discussed between neighborhoods of different poverty levels, and identify differences in topics discussed between neighborhoods of similar poverty levels.

Our goal is to identify segments that denote significant shifts of content (distributions). In turn, this will help to detect differences in deliberation and common interests between advantaged and disadvantaged neighborhoods. This requires us to capture similarities and distinctions between neighborhoods based on: the amount of time neighborhoods with different poverty levels spent discussing the same topics, average similarity in topics discussed between neighborhoods with different poverty levels, and average similarity in topics discussed between neighborhoods with the same poverty levels.

Using the segmentation algorithm we aim to identify segmentations such that segment boundaries indicate qualitative changes in topic distributions. Every neighborhood in the analysis is characterized in this manner and the resulting segmentations are then clustered with a view toward identifying enrichments that hold (or do not) at different poverty levels.
Internet use in communities

This study builds on prior research that explores the relationship between Internet use and local engagement [Hampton and Wellman, 2003, Hampton, 2007, Kavanaugh et al., 2000, Kavanaugh et al., 2007, Kavanaugh et al., 2008, Hampton, 2010]. In particular, we focus on the uneven impacts that Internet use may have on participatory democracy and informal deliberation for communities with a concentration of poverty.

A number of studies have demonstrated that the availability of a relatively simple neighborhood website and discussion forum can increase local tie formation, informal deliberation, and civil and civic behaviors [Hampton, 2007, Hampton and Wellman, 2003, Hampton, 2010]. For example, a longitudinal study of how local social networks changed as a result of a neighborhood email list found that the average person gained over four new local social ties for each year that they used the intervention [Hampton, 2007]. Moreover, the type of discussion that was common in these forums was found to promote collective action and civic engagement [Hampton and Wellman, 2003, Hampton, 2007]. A recent, large, random survey of American adults found that of those who use an online neighborhood discussion forum, 60% know all or most of their neighbors, 79% talk with neighbors in person at least once a month, and 70% had listened to a neighbor’s problems in the previous six months. This compared to the average American, 40% of whom knew their neighbors, 61% talked in-person, and 40% listened to a neighbor’s problems [Hampton et al., 2009].

Characterizing Neighborhoods

We used our segmentation algorithm to track discussions across each individual neighborhood; the next step is to compare such segmentations across neighborhoods.

Recall that since LDA topics are characterized in terms of distributions over terms $p(w|z_n)$ and that such distributions are weighted to yield the joint distribution:

$$p(w, z_n) = p(z_n).p(w|z_n)$$ (4.5)
These distributions (one for each segment of each neighborhood) must now be compared with an aim toward identifying commonalities and discrepancies. However, before we capture distinctions between such distributions, we must ensure that the underlying distributions are expressed over the same vocabulary (terms). To this end, we use the superset of terms from both distributions as the sample space over which two segments induce their respective distributions.

Most clustering algorithms require a symmetric measure of association and we employ the Jensen-Shannon Divergence (JSD):

$$JSD(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M)$$

(4.6)

where

$$M = \frac{1}{2}(P + Q)$$

(4.7)

Note that the Jensen-Shannon divergence is just a symmetrized version of the KL-divergence. The dissimilarity matrix constructed in this manner can be used as input to any clustering algorithm, e.g. an agglomerative clustering with single-linkage criterion is used here.

**Qualitative Methods**

To test our hypothesis, that social media can afford democratic engagement in areas of concentrated poverty, we focus our analysis on where the iNeighbors intervention has been a success. By focusing on the 20 most active iNeighbors groups, previously described in 2.1, we identify local areas that have successfully adopted social media for civic and civil engagement. Traditionally, we would expect to find very few examples of engagement in areas where poverty rates are high—nearly all successful iNeighbors groups should be in areas where there is little concentration of inequality. However, our hypothesis runs counter to this traditional expectation, we expect social media to afford successful democratic engagement in areas where poverty rates are high.
To test our hypothesis that informal deliberation in areas of high poverty would be similar to deliberation that takes place in areas where poverty is low, we modeled how long neighborhoods with different poverty levels spent discussing topics, the average similarity in topics discussed between neighborhoods with different poverty levels, and the average similarity in topics discussed between neighborhoods of similar poverty levels. For the application specific purpose, we used the dataset presented in 2.1. this dataset consists of six neighborhoods, three advantaged and three disadvantaged.

Our goal is to study two basic questions:

- What lengths of time neighborhoods with different poverty levels spend discussing topics?
- What is the average similarity in topics discussed between neighborhoods with different poverty levels, and the average similarity in topics discussed between neighborhoods with similar poverty levels?
Findings

We applied our temporal segmentation algorithm on the six selected neighborhoods. The output of the algorithm is a set of segments from each neighborhood, a dissimilarity matrix, and a dendrogram depicting the clustering of all segments across neighborhoods. Some segments were examined manually, by checking the original text to validate the segmentation output. A partial segmentation output is shown in Fig. 4.3 for a disadvantaged neighborhood and in Fig. 4.2 and Fig. 4.4 for a more advantaged neighborhood.

- Characterizing Segment Durations
Fig. 4.5 depicts the segmentation outputs for the six disadvantaged and advantaged neighborhoods for the one year period in which messages were exchanged within the communities. The segmentation algorithm was applied on each neighborhood separately to identify shifts in topics. Segments identified from each neighborhood are aligned so that vertical ordinates denote the same time point globally. The dashed vertical lines in each segmentation denote the algorithm-picked boundaries. There is not a significant difference in segment durations across the two classes of neighborhoods. The average length of segments from advantaged neighborhoods is 3.24 months, whereas the average length of segments from disadvantaged neighborhoods is 3.38 months. (Note that the segment features a collection of topics during its time, but this does not mean that all these topics were discussed during the entire duration of the segment.)

- Characterizing Topical Content of Segments

We employed our inferred topic models to construct the dissimilarity matrix across neighborhood segments using the approach described earlier. Topics ranged in similarity from 0 to 4.43, where zero means that the two segments are identical.

If discussion topics within disadvantaged neighborhoods were substantively different from
topics within neighborhoods that have lower poverty levels, the divergence coefficient would be significantly higher between advantaged and disadvantaged neighborhoods than it is within neighborhoods that are similar in poverty. That is, we would expect topics within neighborhoods of similar poverty level to be more similar to each other than they are with neighborhoods that are substantively different in poverty.

Across neighborhoods, dissimilarity in segments ranges from 0 (identical) to 4.43, the mean difference is 2.19 ($SD = 1.09$). The mean divergence coefficient between all discussion topic pairs within communities that are low in poverty is 2.18 ($SD = 1.09$), ranging from 0.11 to 4.42. The average divergence between all neighborhoods low in poverty is not significantly different from the average divergence of topics within neighborhoods low in poverty ($M = 2.11, SD=1.01$; one-way ANOVA $>.05$). Topics discussed within low poverty neighborhoods are similar across all low poverty neighborhoods.

The average divergence coefficient between all topic pairs across all high poverty areas ranges from 0.09 to 4.20 with a mean of 2.26 ($SD = 1.09$). Looking within high poverty neighborhoods, the mean divergence is 2.35 ($SD = 1.07$), which is not significantly different...
from the divergence between topics in similar high poverty areas ($M = 2.21, SD = 1.10$; one-way ANOVA $> 0.05$). The variation in topics discussed within high poverty neighborhoods is consistent across high poverty neighborhoods.

Comparing discussion topics in high and low poverty areas, divergence ranges from 0.20 to 4.43 with a mean divergence of 2.16 ($SD = 1.09$). There was no significant difference between divergence within neighborhoods of similar poverty level in comparison to divergence between neighborhoods of contrasting poverty (one-way ANOVA $> 0.05$). Consistent with our hypothesis, the variation in topics discussed within advantaged and disadvantaged areas is not statistically different than the variation in topics between areas of high and low poverty. The range and nature of topics is the same in high poverty areas as was found in more advantaged areas.

A flat clustering of segments reveals congruences as well as outliers. Fig. 4.6 depicts some segments that were clustered together and the topics that contributed to their clustering. Other outliers segments are also shown in the figure. Non-outliers reveal common discussions about topics.

A flat clustering of segments reveals congruences as well as outliers. Non-outliers reveal common discussions about topics.

For example, in Neighborhood 7 [2009-03-01 - 2009-11-01] and Neighborhood 4 [2009-01-01 - 2009-05-01], there were messages discussing the setup of a neighborhood watch meeting and messages discussing a petition. The petition was for commercial vehicles parking in Neighborhood 7 and in Neighborhood 4 it was to save a theater. In Neighborhood 5 [2009-01-01 - 2009-02-01] and Neighborhood 6 [2009-12-04 - 2010-08-04], there were many messages about elementary and middle school events and issues. On the other hand, outliers reveals discussions about an unusual topic. For example, in Neighborhood 3 [2010-01-11 - 2010-09-11], we found a lot of messages discussing a gunshot and a number of burglaries. In this segment, a lot of messages discuss how to buy a gun or a dog. Another example is Neighborhood 4 [2009-05-02 - 2010-01-02], which had an intense discussion after an article
appeared in the local newspaper asking people to vote for either closing a public library or increasing taxes to cover the expenses. The last example is Neighborhood 3 [2010-10-13 - 2010-12-13], which had many messages discussing several dog attacks in the neighborhood, problems with the dog owner, and safety.

Discussion

Here we address the divide in democratic engagement that exists between advantaged and disadvantaged communities. We look for evidence that the gap between high and low poverty communities, in democratic participation and deliberation, is affected by the use of a social media intervention. Specifically, we have argued that new communication technologies afford civic and civil behaviors and informal deliberation in high poverty communities, similar to what is experienced in communities that are low in poverty. Our approach compares the adoption of a new technology across neighborhoods of high and low poverty. We use a unique algorithm to:

- Detect differences in deliberations activity between neighborhoods with different poverty levels.
- Detect whether there are more or less common discussion topics between communities with different poverty levels.

We did not find significant differences between high and low poverty neighborhoods in terms of either the length of discussion periods or the overall topics of discussion. In addition, we found that the rate of adoption of a communication tool for participatory democracy was much higher than would be expected based on established theories pertaining to the digital divide and concentrated inequality. This is not the usual finding in studies of the digital divide, where lower socioeconomic status populations typically have fewer opportunities to participate in public deliberation.

In the past structural constraints internal to disadvantaged communities limited opportunities for deliberation and democratic participation. Social technologies may make communication
possible where it was not before. One possible explanation, as to why social media may be such an important tool for engagement among this population, may relate to the way these technologies bring people together. Previous findings, that use of the Internet as an information tool has a modest positive relationship to engagement for those who are already likely to be engaged [Shah et al., 2005, Boulianne, 2009, Hargittai and Shaw, 2011], do not extend to the truly disadvantaged. However, when the Internet is used as a social tool, a means to communication between people who are “locally” embedded in existing social structures (even if those structures are loosely connected) it affords social cohesion, discussion, and engagement. Technologies that facilitate communication among a population that shares geography, or possibly other sources of affiliation, enables contact that may previously have been desired, but was constrained by physical and structural barriers. It may not be surprising that, when barriers to contact are reduced, we find that residents of high poverty areas are as motivated to participate and deliberate about local issues as people of other communities. If these findings are generalizable, the policy implications are significant. Insuring equal access to social media, across socioeconomic divides, has the potential to reduce persistent inequalities in democratic engagement.

4.2.2 Digging into Historical Newspaper Archives using Dynamic Temporal Segmentations over Topic Models

This work was done as a collaboration with Michelle Seref (Department of English, Virginia Tech), Tom Ewing (Department of History, Virginia Tech), Laura West (Department of History, Virginia Tech), Naren Ramakrishnan (Department of Computer Science, Virginia Tech), Bernice L. Hausman (Department of English, Virginia Tech)

The 1918 influenza epidemic, which killed as many as 50 million people worldwide, has long been recognized as one of the most deadly disease outbreaks in modern world history. This epidemic occurred in the last months of the Great War, which always overshadowed, yet also shaped, discussion of the threat of illness. Because this outbreak occurred at a time when newspapers provided extensive local reports while also communicating national and international news, historians
are interested in understanding how newspaper coverage of the influenza epidemic was shaped by the war-time context, or what we might today call national security threats. For instance, previous research has identified the American and Canadian public’s commitment to war efforts even in the face of serious health threats. Crosby [Crosby, 1989] found that Liberty Loan drives continued to be held in major U.S. cities during October, just as the epidemic was about to take hold. With the greater availability of digitized newspapers, it is of immense interest to analyze ever increasing collections of text archives to shed insight into news coverage and capture important periods in the progression of the epidemic.

One of the key questions that historians would like to answer, as they dig through digital archives, is: what are the key stages of progression in coverage of an issue or phenomenon? Which stage occurred before which other, and do they correspond to known externalities or other factors? Are there critical time points that establish compartmentalization over the full temporal course? Such information, if automatically extracted using analytic techniques, can complement close reading traditions familiar to humanists.

We demonstrate a successful application of our algorithm to archives of the Washington Times. By studying the ebb and flow of ideas in the Fall of 1918 we illustrate how our segmentation algorithm extracts important qualitative features of news coverage of the pandemic.

**How Historians Currently Do Analysis: Perspectives on Rhetorical Research**

Rhetorical and historical research are reiterative practices that spiral through a process that begins with preconceptions, selects data based on these preconceptions, develops new frameworks as a result of data analysis, and moves back to a rethinking of the initial preconceptions based on newly developed frameworks and the knowledge that results from them. Preconceptions comprise prior knowledge, theoretical frameworks, and existing research questions. Prior knowledge on any given topic includes both what we know and what we think we need to learn before we can address the research questions.
Here we use specific examples from our case study described later to elaborate general points. With respect to prior knowledge, for example, we know that the second wave of influenza in the United States passed through the eastern seaboard in the early fall of 1918. We also know that this period coincides with the final months of World War I. Based on this prior knowledge and some initial scanning of various newspapers, we chose to focus on *The Washington Times*, a daily paper in Washington, DC, which at this time published an evening edition. We limited our analysis to September, October, November, and December 1918.

Our research questions concern what is now thought of as “national security interests” but which at the time would have been understood as the “war-time context.” We are interested to know how newspaper coverage of the influenza pandemic was shaped by the war-time context, as well as how coverage of the war was shaped by the influenza threat. These questions send us to the data looking for the impact of context on the reporting of influenza.

Our theoretical frameworks are based in ideological analysis, semiotics (the study of sign systems), narrative analysis, rhetorical analysis, and historical analysis. Ideological analysis address politics, power, and social dynamics, including the analysis of gender, race, and class, and paying attention to vectors of power as they are produced by particular social circumstances. Semiotics studies sign systems, that is the use of words and images to signify particular ideas or frameworks. Semiotics is useful in studying advertising, as well as news journalism more generally. Narrative analysis attends to recurrent themes, repeated word use, typical story lines or plots, and is useful in identifying underlying patterns that are not evident at the literal level of textual content. Rhetorical analysis pays special attention to genre and discourse use in specific situations and contexts. For example, we have identified a number of forensic terms, such as “victim”, “investigate”, and “suspected”, used to refer to influenza in addition to appearing in articles about crime on the front pages of *The Washington Times* during this period. An understanding of the historical context provides the basis for all of these language-oriented interpretations.

At this point in our research, we are only working with context analysis in order to determine the timeline of events which occurred once the flu hit a particular region and became an epidemic.
Doing so allows us to calibrate the manual analysis with the algorithmic elements of the research. An example of context analysis would pay attention to the overall concerns as exhibited on the front pages of papers during this period. Specifically, in every paper in October in which influenza appears on the front page, the banner headline is nevertheless about the war. In addition, October was the last month of the fourth Liberty Loan drive, which was undersubscribed until close to the end of the war. These concerns are interwoven with concerns about the influenza epidemic in Washington, given concerns about crowds and contagion.

In order to read and analyze articles from The Washington Times on influenza during this period, we needed to decide how to select appropriate issues of the newspaper. We did a keyword search in the Chronicling America database (described in detail later) exclusive to The Washington Times between August and December 1918, using the terms, “grip”, “grippe” and “influenza.” A quick scan of the resulting issues determined that most articles of interest were on front pages, so we made an initial decision to exclude non-front-page articles from the analysis. We found that uses of these terms that were not on front pages tended to be advertisements or in articles continued from the front page. We altered our initial decision to include August when we discovered that there was only one instance of the use of “influenza” in that month, and it was in an advertisement.

Historical and rhetorical analysis depends on close reading of data. When we read, we look for patterns (i.e. repetition) of word use, topics, and themes. In rhetoric, this practice can be systematically applied as “coding.” We look for both expected and unexpected patterns of usage. Our expectations are based on prior knowledge and our theoretical frameworks, which tell us what we think we will find. Thus when we find such information in our data, we note it. However, we also pay attention to findings that we do not expect – what seems unusual or contradictory to what we think we know. Unexpected findings might be words used that we didn’t think to search for, an example might be, “flu”, which in these articles and titles, is always placed within quotation marks. We are still not sure what to make of this finding. We also look for the placement of articles on the page. In addition, we often have to conduct new research to make sense of findings whose meaning is not entirely clear to us. For example, we are currently investigating the extent of newspaper censorship during this period, given that most of the coverage of the flu in October 1918 seems to
be very local to Washington.

To analyze our findings once they have been determined from the data, we use theory and prior knowledge as frameworks to narrate our explanations. Analysis must account for both the expected and the contradictory or new information. Analysis creates new narratives bringing latent elements from the data to the level of manifest content. Rhetorical analysis pays special attention to the contexts of discourse and the influence of context on reception, understanding, and use. How do people use the discourses available to them to make arguments, explain things, or justify themselves? What is the purpose of specific forms of discourse use and are they successful or not? How do unintended meanings (ideology) make their way into utterances and written discourse and what are their modes of circulation and influence? These are the questions we seek to answer using the segmentation algorithm.

In this work we used the Chronicling America Dataset 2.2. Two projections (sub-datasets) from this collection to apply the segmentation algorithm on were created. First projection is The Washington Times front pages and the second is Influenza paragraphs Extracted from the same newspaper. The focus of this work was on The Washington Times for the period from September 1918 to December 1918.

To decide whether our segmentation approach reveals important insights we compared it’s output with a manual analysis conducted by a group of three historians. The goals of the study was to understand the event timeline and obtain a conceptual understanding of the coverage of influenza in The Washington Times during this period. The manual analysis steps involved identifying the sequence of events in Washington following the outbreak of flu in late September through the end of the epidemic in late October. We follow the discussion of influenza, policies to close schools, theaters, and churches, and other public health decisions in the city. Our goal is to see if topic modeling and segmentation can provide results that mirror analysts’s manual traditional analysis of the papers – i.e., actual reading and interpretation. The event timeline created from manual rhetorical and historical analysis thus far is shown in Table 4.1
Table 4.1: Event Timeline created from Front Pages of The Washington Times (1918).

<table>
<thead>
<tr>
<th>September 1918</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sept 14</td>
</tr>
<tr>
<td>Sept 15</td>
</tr>
<tr>
<td>Sept 18</td>
</tr>
<tr>
<td>Sept 19</td>
</tr>
<tr>
<td>Sept 21</td>
</tr>
<tr>
<td>Sept 22</td>
</tr>
<tr>
<td>Sept 24</td>
</tr>
<tr>
<td>Sept 26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>October 1918</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oct 1</td>
</tr>
<tr>
<td>Oct 2</td>
</tr>
<tr>
<td>Oct 3</td>
</tr>
<tr>
<td>Oct 4</td>
</tr>
<tr>
<td>Oct 5</td>
</tr>
<tr>
<td>Oct 6</td>
</tr>
<tr>
<td>Oct 7</td>
</tr>
<tr>
<td>Oct 8</td>
</tr>
<tr>
<td>Oct 9</td>
</tr>
<tr>
<td>Oct 10</td>
</tr>
<tr>
<td>Oct 11</td>
</tr>
<tr>
<td>Oct 12</td>
</tr>
<tr>
<td>Oct 13</td>
</tr>
<tr>
<td>Oct 14</td>
</tr>
<tr>
<td>Oct 15</td>
</tr>
<tr>
<td>Oct 16</td>
</tr>
<tr>
<td>Oct 17</td>
</tr>
<tr>
<td>Oct 18</td>
</tr>
<tr>
<td>Oct 19</td>
</tr>
<tr>
<td>Oct 20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>November 1918</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov 3</td>
</tr>
<tr>
<td>Nov 7</td>
</tr>
<tr>
<td>Nov 16</td>
</tr>
<tr>
<td>Nov 21</td>
</tr>
<tr>
<td>Nov 30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>December 1918</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec 6</td>
</tr>
<tr>
<td>Dec 8</td>
</tr>
<tr>
<td>Dec 9</td>
</tr>
<tr>
<td>Dec 10</td>
</tr>
<tr>
<td>Dec 12</td>
</tr>
<tr>
<td>Dec 19</td>
</tr>
<tr>
<td>Dec 20</td>
</tr>
</tbody>
</table>

Findings

We now outline below comparisons between the segmentations discovered by the segmentation algorithm and their relationship to the manual analysis.

- Modeling front pages of The Washington Times

The 1918 September through December topic modeling with segmentation of front pages of
Figure 4.7: Segmentation results for The Washington Times Influenza paragraphs from September 1918 to December 1918.

Figure 4.8: Segmentation results for The Washington Times front pages from September 1918 to December 1918.

The Washington Times demonstrates that the war is the main contextualizing topic throughout this entire period. Segmentation output is shown in Fig. 4.7. The segmentation at 9/22 and 10/21 roughly corresponds with the outbreak of the flu in Washington and its most virulent epidemic period. The period from 10/21 to 11/05 corresponds with the intense negotiations about the armistice. The war was reported to end on 11/7, but in actuality the full armistice occurred on 11/11.

- Modeling paragraphs with Influenza

The segmented topic modeling of influenza paragraphs on the front pages of The Washington Times track the influenza epidemic well. Segmentation output is shown in Fig. 4.8. The first
segment, 9/16-9/23, concerns outbreaks in other cities and the first outbreak of influenza in Washington on 9/20. The second segment, 9/23-10/22, concerns the epidemic as it develops until its peak and then waning, which occurred 10/22. The following segments (10/22-10/30 and 10/30-11/28) demonstrate the waning epidemic and its aftermath, including reopening of schools, churches, and theaters. By the last segment, 11/28-12/20, the flu is not really a viable topic, as those topic clouds that were developed are simplistic, contain a small number of terms, and are largely the same.

The topic clouds in the first section name the first victim (Henne), although they do not include place names (like New York and Chicago) that were mentioned in articles. Four of the five topic clouds have the word, ‘Spanish,’ in them, referring to the, ‘Spanish Influenza.’ We see indication of the first reported cases in Washington, DC, in another cloud.

The month-long segment from 9/23 to 10/22 covers the main period of influenza epidemic in Washington, DC, and includes the following clearly delineated topics: (1) closing of schools, churches, and other public meeting places, as well as government offices, (2) health care facilities, personnel, and treatment, (3) economic concerns of the war in conjunction with influenza, i.e., the Liberty Loan drives, and (4) reporting that describes the number of cases in the district, usually in a 24-hour period. There is one cloud with an unclear topic, - it seems to include preventive measures but also indicates more global concerns (mention of Germans and the world). We think that this topic cloud may not really be a unique or separate topic. Also, in this segment we did not notice a discussion of, “masks,” although there is a discussion of gauze masking in the newspaper. Indeed, the word, “mask,” does not appear in any of the clouds.

The next segment is only a week long, from 10/23 to 10/30, and seems to correspond with the immediate waning of the epidemic and the reopening of closed schools, churches and theaters. We can detect five separate topics in the clouds, yet it was harder to determine these than in the previous segment. The newspaper reported the epidemic receding on October 21, with a sudden jump in deaths the following day, which were nevertheless interpreted as not indicating a change in the reduction of the illness’s spread. Reporting on October 23
indicates that the epidemic is receding and at the very end of the month the public buildings and meeting places are to be reopened. The topics we identified include (1) time, which we think references the future reopening of schools, churches, and theaters, (2) sports and leisure, with some mention of army camps, (3) government and administration, which may have to do with the return of normal working hours, (4) reporting of cases and the continued reopening of public venues, and (5) a general discussion of flu in the city.

In the segment from 10/30 to 11/28 it is very difficult to determine separate and unique topics for the clouds. There does not seem to be either commonality or distinction. We think that there may not really be five topics, but actually perhaps just one or two. Preventative themes are prevalent in at least two clouds. There also seems to be mention of things starting up again. The following segment, from 11/28 to 12/20, really does not seem to have more than one conversation, if that, as the clouds are minimal and uninformative. The segment following, 12/20-12/30, only has two words “year and influenza,” in the entire grouping of five clouds, two of which have only one word (year). These diminishing topic modeling cloud groupings indicate that the flu is truly becoming less of a topic of conversation on the front page of The Washington Times during this period.

**Discussion**

Our results demonstrate that using the segmentation algorithm enabled us to clearly follow the rise, height, and fall of the influenza epidemic in Washington, DC, as reported in The Washington Times throughout the fall of 1918. The segmentation strategy appears to be most successful in capturing conversations during the period in which there was the most reporting on the epidemic, i.e., the period of its greatest virulence and spread in the city, 9/23-10/22. Before and after this period, the topic modeling clouds are more difficult to interpret, but the segmentation does seem to clearly follow events.
4.3  Summary

In this chapter, we presented a time series segmentation algorithm that segment time based on shifts in topics. We applied the algorithm on two different datasets: Historical newspapers dataset and i-Neighbors dataset. The algorithm served different purposes in these applications.

In the digging into Historical Newspaper archives application the goal was to understand the progression in coverage on the 1918 influenza from The Washington Times newspaper. In this application, the algorithm was successful in detecting the ebb and flow of ideas and in extracting the pandemic qualitative features of news coverage.

In the i-Neighbors application, our goal was to study the conversation patterns in advantaged and disadvantaged communities and how new communication technologies changed the behaviors and informal deliberation across neighborhoods. The algorithm was successful in capturing the similarities between neighborhoods and the time duration in which they spent on discussing topics.

Through these two applications, we showed that the algorithm was a great assistance to experts working on these projects. We approached the evaluation of the segmentation algorithm from a qualitative perspective by experts closely examining of the algorithm output.
Chapter 5

New Visual Analytic Representations

Data mining algorithms have evolved greatly over the past years, especially for topical text modeling [Blei et al., 2003]; however, capturing key breakpoints in topic evolution and defining appropriate visual representations for such breakpoints is an understudied problem.

On the other hand, one of the most frequently used visualization tools for topics is Tag clouds. Tag clouds are visualizations for keyword groupings (topics), where the size of each keyword represents its relative importance within a topic. Unlike ThemeDelta, there is no visual method to track topic evolution and the scattering and gathering of keywords. Others, like ThemeRiver [Havre et al., 2002] use Streamgraphs to visualize topics across time. Streamgraphs does not show the scattering and gathering of keywords into topics (trends), however, is capable of conveying the structural features of the corpus. Although much research exists in data mining and visualization, we posit that they are insufficient to address the needs of these emerging applications.

ThemeDelta, a new temporal topic modeling approach, will be presented here. The main difference between ThemeDelta and existing approaches is that it can automatically identify segments where significant topic shifts occur. To capture topic shifts, we embed a temporal segmentation algorithm around a topic modeling algorithm, as discussed in the previous chapter. We use the width of each trend line to communicate the prominence of its trend in the dataset at a particular
ThemeDelta is intended to convey local and global temporal changes in the distribution of evolving trends. The system detects and visualizes how different trends converge and diverge into groupings at different points in time, as well as how they appear and disappear during a time period. The

Figure 5.1: ThemeDelta visualization for Barack Obama campaign speeches during the U.S. 2012 presidential election (until September 10, 2012). Green lines are shared terms between Obama and Romney. Data from the “The American Presidency Project” at UCSB (http://www.presidency.ucsb.edu/).

5.1 ThemeDelta Overview

ThemeDelta is intended to convey local and global temporal changes in the distribution of evolving trends. The system detects and visualizes how different trends converge and diverge into groupings at different points in time, as well as how they appear and disappear during a time period. The
system consists of two major components: a backend data analytics component, and a frontend visualization component:

- The **analytics backend** is responsible for accepting a large temporal text corpus and automatically identifying segments that characterize significant shifts of coverage. The algorithm responsible for this task was originally developed to detect deliberation in social messages [Gad et al., 2012] and presented in the previous chapter.

- The **visualization frontend** is responsible for graphically representing the discovered trending of topics. While originally designed for timestamped text collections, we see many additional applications such as for genealogy (e.g., [Kim et al., 2010]), communication graphs (e.g., [Elmqvist and Tsigas, 2003]), and general dynamic graphs.

### 5.1.1 Data Format

The backend accepts a text dataset consisting of timestamped data. The frontend takes the output of the backend for visualization. The backend output consists of **trends**, **topics** (groups of trends at a specific point in time), and **segments** (a closed interval of time, modeled as a group of topics).

The exact mapping of these general concepts to a dataset is domain-specific. For example, for a timestamped document collection, trends could represent the terms extracted from the documents, and the topics would model how these terms converge into groupings at different points in time.

### 5.1.2 Implementation

ThemeDelta is a web-based application that is built to be capable of running in any modern web browser. To realize this the backend of ThemeDelta was built using Java and this include all data preprocessing, clustering, and segmentation.

The frontend was built using JavaScript and SVG. The current implementation is fully
interactive and animated, and is built using the Raphaël\(^1\) toolkit for scalable vector graphics.

![Diagram of ThemeDelta visual representation](image)

**Figure 5.2: Basic visual representation used by ThemeDelta.**

### 5.2 ThemeDelta: Visual Representation

ThemeDelta’s visual representation draws on TextFlow [Cui et al., 2011], and uses a basic visual encoding consisting of sinuous *trendlines*—each representing a trend in the dataset—stretching from left to right along a timeline mapped to the horizontal axis (Figure 5.2). The horizontal space along this axis is divided equally among different time segments \((t_1, t_2, \text{ and } t_3\) in Figure 5.2). Topics for each segment are perceptually conveyed by clustering the trendlines for the grouped trends next to each other along the vertical axes, leaving a fixed amount of empty vertical space between adjacent topics. Vertical lines, one for each time segment, partition their horizontal positions.

### 5.2.1 Visual Design

Given this basic design, many design parameters remain open. Below we review the most important of these and motivate our decisions for the visualization technique. A visualization developer

\(^1\)http://raphaeljs.com/
using the same basic visual representation may make different choices than these depending on the application.

**Shape.** To communicate the organic nature of evolving trends, we use splines to yield smooth curves. The resulting lines are continuous, predictable, and appealing. An alternative design would have used rectilinear or sharp angles, but curves are likely easier to perceive and more aesthetically pleasing.

**Thickness.** Trendline thickness is a free visual variable. While it is possible to use a uniform thickness for all trendlines, it can also be used to convey scalar data for each time segment. Because increasing thickness will raise the visual salience of a trendline, we tend to use it to convey the weight of each keyword calculated by our segmentation algorithm.

Furthermore, our visual representation uses vertical dashed lines to partition time segments on the visual space. The thickness of these lines is another free variable that can, e.g., be used to indicate the relative extent of each time segment. This is useful since time segments may be irregular; some segments are significantly longer than others.

**Color.** Color is another free parameter in our visual representation, and can convey either a quantity (using a color scale) or a category (using discrete colors). The choice depends on the application. For example, we use it both to show the strength of a correlation, as well as to convey which entity class a particular trendline belongs to.

**Discontinuities.** Trendlines can begin and end at any time segment, sometimes only to reappear later in time. We communicate this using a tapered endpoint of the line (see borders in Figure 5.2). An alternative design could have dashed the trendlines for the periods of time where there is no associated value, similar to the use of different trend shapes in TextFlow [Cui et al., 2011]. We chose to avoid this to minimize visual complexity.

**Labels.** We draw the names of each trendline on the line itself for each time segment. While this is redundant (one instance of the label is sufficient) and potentially a source of visual clutter, it prevents the user from having to trace an undulating trendline back to a single label at the far end of
the visualization. We also scale the label size based on the trendline’s thickness, similar to word
scaling in word clouds.

**Duplicated Trends.** Sometimes a trend may exist in more than one topic for a particular
time segment (see trend A at time $t_2$ in Figure 5.2). To make the visualization consistent, as well as
to convey the fan-out, we are forced to fork the trendline into two or more pieces. Analogously, in a
time segment following a duplicated trend instance, the trendlines should be merged to maintain
consistency. In situations when there is more than one candidate to fork from or merge to, we
choose the two trend instances that are vertically closest to each other (see the layout algorithm
discussed below).

### 5.2.2 Interaction

Several interaction techniques are meaningful for ThemeDelta frontend. First of all, geometric
zoom and pan allows for being able to magnify a certain part of the visualization to see details.
Furthermore, hovering over a trendline will highlight the line, including all of its branches in other
parts of the visualization (even past a discontinuity). Figure 5.1 shows this interaction, where
the trend lines associated with the keyword *energy* are highlighted in response to a mouse hover
interaction.

The interface also supports searching for trendlines by name. In addition, we provide a
combined filtering and resorting operation. Clicking on a trendline will add it to a filter box, causing
the layout to be recomputed with the selected trendline at the top of the screen. The new layout
will only include trendlines that are connected to the selected trendline, i.e., which in at least one
time segment belong to the same topic as the selected trendline. Following the example presented
in Figure 5.1, clicking on the trend line for the keyword *energy* performs the filtering operation,
and the visual layout is changed such that the filter keyword is positioned at the top (Figure 5.3).
Additional trendlines can be added to the filter box, yielding a conjunctive filter (only trendlines
which are connected to all selected trendlines are shown).
5.2.3 Layout

The ThemeDelta frontend visualization layout divides the available horizontal space equally between time segments, while vertical space is divided locally between the topics associated with each time segment. Due to the ever-changing topic groupings over time as well as the dynamic appearance and disappearance of trends, it is typically not possible to represent trends as straight lines. In fact, a single trend could appear in a different topic and at a different vertical position with each new time segment. This is the reason for using smooth splines to convey this organic trend evolution.

Of course, this in turn means that trendlines will frequently cross one another while connecting the multiple occurrences of a single term across different time segments. Research in graph drawing has shown that the ease with which a user can follow an edge depends on the number of crossings with other lines in its path [DiBattista et al., 1998].

Tanahashi and Ma [Tanahashi and Ma, 2012] discussed a set of layout design principles for better legibility of storyline visualizations like ThemeDelta. However, the complexity of their algorithm makes it difficult to achieve real-time layout updates. Other work proposed by Liu et al. [Liu et al., 2013] trade-off optimal layout with algorithm performance to achieve real-time updates. The algorithm used for ThemDelta is similar to the one proposed by Liu et al. [Liu et al., 2013]. However, contrary to their algorithm, we do not have hierarchical relationships in the underlying data and to facilitate the identification of individual topics by supporting a constant
reasonable vertical space between them, the layout algorithm used in this chapter does not perform
the topic alignment step. Moreover, to achieve real time interactivity our implementation minimizes
line crossings through a single iteration across different time segments.

In particular, ThemeDelta relies on a deterministic layout algorithm that minimizes trendline
crossings by first sorting the vertical positioning of different topics, followed by sorting the trends
within each topic. While sorting topics at a particular time segment $t_i$, a topic $p_1$ is placed before
another topic $p_2$ if the average vertical position of the trends contained in topic $p_1$ is less than the
terms present in topic $p_2$ at the previous time segment $t_{i-1}$. Topic position in the first time segment
is either determined randomly, or using some attribute of the underlying data.

After sorting topics it is time to sort the trends within each topic. Except for the first time
segment, trends within a topic are sorted such that their relative vertical position remains the same
as it was in the previous time segment. Once all trends are sorted, the trends contained within topics
of the first time segment are sorted such that their vertical position remains the same as in the second
time segment.

Figure 5.4: Comparison of different stages of the layout sorting algorithm used for the ThemeDelta

Figure 5.4 shows the progressive decrease in the number of trendline crossings at different
stages of the layout. In Figure 5.4(a) the dataset is visualized without any sorting. This results in
a total of twelve crossings between trendlines, connecting multiple occurrences of terms across
the two time segments $t_1$ and $t_2$. Figure 5.4(b) shows the resulting layout after topic sorting. As
shown in the figure, the topics within time segment $t_2$ are now ordered based on the average vertical
position of their corresponding terms within time segment $t_1$. This ordering of topics has reduced the number of line crossings from 12 to 6. Finally, Figure 5.4(c) shows the resulting layout after trend sorting. Here again it is evident that the number of line crossings is reduced even further. All in all, as a result of the layout algorithm, the number of trendline crossings has been reduced from 12 to 2.

Figure 5.5: ThemeDelta visualization for Mitt Romney campaign speeches for the U.S. 2012 presidential election (as of September 10, 2012). Green lines are shared terms between Obama and Romney speeches. Data from the American Presidency Project at UCSB (http://www.presidency.ucsb.edu/).

## 5.3 Domain Specific Applications

### 5.3.1 U.S. 2012 Presidential Campaign

Political speeches, especially during an election campaign, are particularly interesting document collections to analyze because the political discourse tends to change and evolve as different candidates respond and challenge each other over the course of the campaign. Visualizing the speeches of different candidates would allow for comparing the trends of each candidate with each
other. To study such effects, we used the U.S. 2012 presidential election campaign speeches.

The U.S. presidential election takes place every four years (starting in 1792) in November (the 2012 election day was November 6), and is an indirect vote on members of the U.S. Electoral College, who then directly elect the president and vice president. In 2012, the Republican and Democratic (the two dominant parties, representing conservative vs. liberal agendas) conventions were held on the weeks of August 27 and September 3, respectively. The two opposing candidates were Republican nominee Mitt Romney, and Democratic nominee Barack Obama (incumbent President of the United States). The ThemeDelta for both candidates is shown in Figures 5.1 and 5.5.

In collecting data for the United States presidential election, we used campaign speech transcripts for both candidates, first presented in 2.3. For Mitt Romney, we used transcripts from 46 speeches over a 62-week period: from announcing candidacy on July 29, 2011, to August 14, 2012. This corpus included speeches from both the Republican primary election (settled on May 14, 2012 as the main competing nominee Ron Paul withdrew). For Barack Obama, we used transcripts from 40 speeches over a 44-week period: November 7, 2011 to September 17, 2012.

Visualizations of the two candidates Barack Obama and Mitt Romney are shown in Figure 5.1 and Figure 5.5. Trendlines in both visualizations represent characteristic keywords that each candidate uses as a theme in his speeches. Democratic trendlines are colored blue, Republican ones are red, and trendlines for keywords that both candidates share are green.

For the Romney dataset (Figure 5.5), there is a clear impact of time on keywords and topics that the candidate is using. Romney’s message starts out relatively simple with only two main topics, but quickly branches out in complexity as time evolves. The effect of main competitor Ron Paul withdrawing in May is clear: before this date, Romney is trying to win the party nomination, whereas afterwards, he is going for the presidential seat. As a result, his message becomes more simple again: both the number of keywords and the number of topics decreases during the last three segments, presumably to focus on key issues in the Republican election platform.

For the Obama dataset (Figure 5.1), a good portion of the identified keywords are common
with Mitt Romney (i.e., green in color). This could be seen as Obama discussing many of the issues that has become central to the U.S. presidential race. Furthermore, there is a clear presence of keywords such as “health,” “insurance,” and “care,” which may refer to the president’s health care reform from 2010 (informally called Obamacare). This is a controversial issue that still causes a major divide between voters; a Reuters-Ipsos poll in June 2012 indicated that a full 56% of Americans were against the law.

Taken as a whole, both datasets have a heavy emphasis on economics keywords. This is commensurate with the overall theme of the 2012 presidential race, which largely has focused on the poor economic situation of the United States.

5.3.2 i-Neighbors Social Messages

The Internet facilitates informal deliberation as well as civic and civil engagement. Web-based applications for informal deliberation (e.g., i-Neighbors [iNe, 2012]) facilitate the collection of data that we can analyze to provide insight into how neighborhoods with different poverty levels use ICTs for informal deliberation. Using ThemeDelta, we can characterize differences and detect common interests in informal deliberation between advantaged and disadvantaged communities.

The goal of this application was to study two basic questions: what lengths of time neighborhoods with different poverty levels spend discussing topics? And what is the average similarity in topics discussed between neighborhoods with different poverty levels, and the similarity in topics discussed between neighborhoods with similar poverty levels?

The data for this application was collected through the i-Neighbors system, first presented in 2.1. When we collected the data in 2010, the i-Neighbors website had over 100,000 users who had registered more than 15,000 neighborhoods. Over 1,000 neighborhoods were active with more than 7,000 unique messages contributing to neighborhood discussion forums. We collected data from six geographically diverse communities located in Georgia, Maryland, New York, and Ohio. We selected the three groups located in areas with concentrated levels of poverty (a poverty rate of
25% or more, 2009 American Community Survey, US Census Bureau) who exchanged the most messages, and the three most active groups in more advantaged areas.

![Diagram of word co-occurrences over time]

Figure 5.6: Result of searching for the word “watch” in low-poverty neighborhood.

We applied our temporal segmentation algorithm on the six selected neighborhoods. Topics within each segments can be examined using the visualization to find topic similarities between neighborhoods. Segmentation labels indicating segments size can be used for comparing the time spent by different neighborhood discussing certain topics.

A partial segmentation output is shown for a disadvantaged neighborhood in Figure 5.7 and Figure 5.8 for a more advantaged neighborhood. From these two examples, the segments sizes are not very different and we can conclude that both the disadvantaged and advantaged neighborhoods spend similar amounts of time discussing topics.

Examining the words groupings in both neighborhoods can lead to discovering differences and similarities in their discussions. For example, in the low-poverty neighborhood in segment [Feb 1, 2009 to Nov 1, 2009], there is a topic that has the words “watch” and “neighbor,” which lead us to conclude that there were some arrangements or discussions about a neighborhood watch. This topic is not found in the the disadvantaged neighborhood visualization. If the user searched for the word “watch” this will result (Figure 5.6) in only showing the topics that has the this word and any other related topic.

Similarly, an example of similarities of topics discussed between neighborhoods can be
shown by examining the segment [Jan 03, 2010 to Sept 3, 2010] in the advantaged neighborhood (Figure 5.8) and segment [Feb 4, 2010 to Oct 4, 2010] in the disadvantaged neighborhood (Figure 5.7). In both segments, there exist two topics in which both communities discuss a park-related project.

5.3.3 Historical U.S. Newspapers

Newspaper stories are precisely the type of ongoing, evolving trend datasets for which ThemeDelta was designed. Below we review the source, segmentation, and visualization for a dataset consisting of historical U.S. newspaper stories from 1918.

Our data source was a historical newspapers database, first presented in 2.2. Some of
newspapers included in this example are: *The Washington Times* (Washington, DC), *Evening Public Ledger* (Philadelphia, PA), *The Evening Missourian* (Columbia, MO), *El Paso Herald* (El Paso, TX), and *The Holt County Sentinel* (Oregon, MO). We gathered data from them, restricting the time to the period September 1918 through December 1918. From this dataset, we extracted only paragraphs that mention the word “influenza” resulting in 2,944 paragraphs. This corresponds to the 1918 flu pandemic (also known as the “Spanish flu”) which spread around the world from January 1918 to December 1920, resulting in some 50 million deaths.

Applying the dataset to ThemeDelta using a weekly segment granularity yields four discrete time segments over the four-month time period. Figure 5.9 shows a visualization of the result, where the transparency value of each trendline has been mapped to the global ranking of the keyword corresponding to the trendline. The thickness of the trendline conveys the ranking of each keyword for a particular time segment, calculated by our segmentation algorithm.
Figure 5.9: ThemeDelta visualization for newspaper paragraphs during the period September to December in 1918. Color transparency for different trendlines signify the global frequency for that keyword.

Figure 5.9 offers several observations that summarize the qualitative nature of trends exposed by ThemeDelta. The output is showing many events that were related to the 1918 pandemic in the data. For example, in the first time segment, September 9 until October 9, there are a topic that contain the terms “mask” and “German.” This corresponds to advisories and guidelines recommending people to use masks to protect themselves from the ongoing influenza pandemic during World War I. In the same segment, the words “liberty,” “loan,” and “campaign” appeared in one of the topics, and continued appearing in the following segment, October 10 until December 5, because a liberty loan campaign were issued to support the army during World War I. Also, in the October 10 to December 5 segment, the army men left the camps to go back home from service and stay with their families; this explains the topic with the words “family,” “home,” “serves,” “spent,” “wife,” and “son.” This topic appeared along with the topic with the words “case,” “disease,” “mask,”
“cross,” and “red” because the returning soldiers were exposed to the disease and some of them were sick. As a result, families were advised to take protective measures.

World War I ended on November 11, 1918, which explains the disappearance of the word “German,” but the country continued suffering from the disease. The word “mask” reappeared back along with “epidemic,” “hospital,” and “disease” in the December 14 until December 28 segment, which aligns with the second influenza wave. Again, during this time people were advised to wear masks to slow down the spread of the disease. The Red Cross was frequently mentioned in the last three segments, which is indicative of the second, deadlier wave of the pandemic that began in October. In both the December 6 to December 13 and December 14 to December 28 segments, the terms “people,” “ban,” and “meet” appeared because people were banned from meeting each other as a precaution measure to limit the spread of the disease. The term “president” appeared in the last segment along with “service” appeared initially in the first segment and then returned with significant strength in the last segment, illustrating the seriousness accorded to the national scale of the pandemic.

5.4 Qualitative User Study

To validate the utility of the ThemeDelta system, including both its temporal segmentation algorithm as well as its visual representation, we conducted a qualitative user study involving expert participants. The purpose was to study the suitability of the approach for in-depth expert analysis of dynamic text corpora. Because of our existing collaboration with historians (the sixth author of this work is a historian), we opted to use the historical U.S. newspaper dataset and engage experts from the history department at one of our home universities.

We used historical data from five U.S. newspapers for our qualitative evaluation from three different areas: New York, Washington, D.C., and Philadelphia. The data was collected from the Chronicling America website\(^2\) and focused on the 1918 influenza epidemic, which killed as many

\(^2\)http://chroniclingamerica.loc.gov/
as 50 million people worldwide and has long been recognized as one of the most deadly disease outbreaks in modern world history. Historians are interested in reconstructing the timeline of events, with a view to understanding previously concealed or neglected connections between public opinion, health alerts, and prevailing medical knowledge.

5.4.1 Method

We recruited three graduate students as participants: one from the history department and two from the English department at our university. The participants were all required to have prior knowledge of America around the Great War/First World War period. Two participants were Ph.D. students and one was a Masters student. We required no particular technical skill prior to participation. While the number of study participants may appear to be low, we want to emphasize that these participants represent a highly expert population and that our study protocol is focused more on an expert review [Tory and Möller, 2005] rather than a comparative or performance-based user study.

The total study time was an hour. The procedure was as follows: Participants were first asked to fill out a background questionnaire. Then the study moderator explained the tool and its features, followed by the task the participants were asked to perform using the tool. After that, the participants were asked to solve several high-level tasks (reviewed below) using the tool. Finally, they were asked to complete a post-session questionnaire to collect feedback on the tool.

The tasks that we asked the participant to accomplish with the help of our system was answering some questions on the 1918 influenza pandemic. Participants were encouraged to refer to the visualization in their answers by mentioning segments names, giving examples, or taking screen captures from the visualization. Tasks were divided into change and connection questions, to allow us to determine whether the visualization and algorithmic choices we made were helpful or not. The change-focused questions were:

- How did the newspapers describe the spread of influenza?
- How does the description of the pandemic change over time?
• Are there different times when the influenza pandemic becomes less important? What are those time periods?

Questions that were focused on connections were:

• What are the categories that appear to be associated with influenza in different newspapers?

• Was there a specific feeling that surrounded the influenza reporting in the newspapers?

### 5.4.2 Results

All three participants were successful in accomplishing the task using ThemeDelta. We determined this by comparing their answers to the task questions with model answers provided by the history faculty collaborator (reviewed in Section 5.3.3). They correctly reported the sentiments that surrounded the influenza from the five newspapers. They also successfully described the change in reporting of the influenza spread. Finally, they all succeeded in discovering the connection between influenza and other categories (e.g., schools, war, and hospitals).

The subjective results of the study were overall positive and the participants all vouched for the helpfulness of the system and the need for such systems in their research. None of the participants had previous experience using any visual analytics systems. This implies that the participants found ThemeDelta to be understandable and easy to use.

All the three participants finished the tasks within the allocated time. They also uniformly reported that the same type of task, if done manually as part of their own research, would normally take several days if not weeks. This highlights an additional strength to our system: minimizing the time spent on manual analysis of large amounts of text, allowing the analyst to focus on collecting insight instead.

In the post-session questionnaire, participants were asked to give their feedback on specific ThemeDelta features. The features that were reported as very useful were labels, line thickness,
duplicate trends, and discontinuations. Participant ratings for other features ranged from very useful to not useful at all, the latter typically because they did not use that particular feature. Some of the identified weaknesses of the tool included not being able to see full phrases or word combinations, managing keyword filtering, controlling the dynamic layout, and high complexity for large datasets.

5.5 Summary

We presented ThemeDelta a visual analytics system we built to help detect the scatter and gather of trends in text corpora. We used the system for three different scenarios; each had its dataset. Datasets used in the scenarios were historical newspaper dataset, presidential campaign dataset, and i-Neighbors dataset.

First scenario was historical U.S. newspaper Spanish flu pandemic coverage. Here we were focused on how newspapers in year 1918 discussed the second wave of the pandemic topic and how these topics temporally evolved. Second scenario was Barack Obama and Mitt Romney U.S. 2012 presidential campaigns. In this scenario, our focus was to identify the similarities between the two candidates and how the topics they discussed in their campaigns evolved over time. Third and last scenario was social messages exchanged between virtual communities via the i-Neighbors. The focus here was on comparing advantaged and disadvantaged neighborhoods from the topics, and the time duration spent on topics perspectives.

The system showed great success in identifying trends and their temporal evolution in the three scenarios. We qualitatively evaluated the system by running an expert user study. The study results showed how successful the system was in helping experts reach conclusions and identify key trends.
Chapter 6

Dynamic Spatial Topic Model

The main goal of this chapter is to extend the basic topic model to accommodate location and temporal distinctions in large document sets. In this chapter, we present a new dynamic spatial topic model (DSTM), a true spatio-temporal model. DSTM can model relationships between locations, topics, documents, and terms in a dynamic fashion. The model enables summarizing and navigating unstructured time stamped text documents while capturing the evolution of topics along with location distribution over these topics.

Previous work in Temporal topic models by [Blei and Lafferty, 2006, Wang and McCallum, 2006, AlSumait et al., 2008, Gohr et al., 2009, Zhang et al., 2010, Hoffman et al., 2010, Hong et al., 2011] and in Spatial topic models [Pan and Mitra, 2011, Wang et al., 2009] do not model the decomposition of topic models into specific topics for specific locations over time. Tracking the evolution of topics and their location overtime is a critical step toward understating major events such as an epidemic or an unrest.

The DSTM model assumes words in a document are reliant on both topic distributions and location distributions. Unlike LDA, this model results in topics distribution over the vocabulary and location distribution across all topics and the evolution of topics and their locations are captured over time. Our model inherits some features from both Author-Topic Model previously proposed
by [Rosen-Zvi et al., 2004] and Dynamic Topic Model previously proposed by [Blei and Lafferty, 2006]. One of the advantages of our model over these two models is that it companies the power of both. We applied the algorithm on multiple newspapers from the Chronicling America repository introduced in 2.2 to understand the differences between those papers in the coverage of the flu as it spread.

### 6.1 Proposed Model

Here we propose a dynamic spacial topic model (DSTM) that incorporate reporting locations into the process of inferring topics. Fig. 6.1 presents our proposed model for modeling time-stamped data. A (Dirichlet) distribution over topics is first organized and, concomitantly, a (Dirichlet) distribution over locations is organized. Next, a (multinomial) topic distribution and a (multinomial) location distribution are picked. The first to incorporate information about a document in the topic inference was [Rosen-Zvi et al., 2004]. Finally, we select a word from the topic distribution and location from the location distribution. Specific model notation is given in Table 6.1.

In order to capture the evolution of topics and locations over time, we assume that $\phi_t$ and $\lambda_t$ are Dirichlet distributions that evolve by adding white (Gaussian) noise at each time step to the distributions resulting from the previous time slice as in [Blei and Lafferty, 2006]. This is done by chaining $\phi_t$ and $\lambda_t$:

$$\phi_{t,k} | \phi_{t-1,k} \sim Dir(\phi_{t-1}) + N(\mu, \delta^2)$$

where $N(\mu, \delta^2)$ reflects the added gaussian noise.

The generative process for time slice $t$ of a chronologically ordered time stamped documents in a corpus is as follows:

1. Randomly draw $K$ multinomial distributions from $\phi_t$, where $\phi_{t,k} | \phi_{t-1,k} \sim Dir(\phi_{t-1}) + N(\mu, \beta^2)$. 
Table 6.1: DSTM notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>number of words in a document.</td>
</tr>
<tr>
<td>$D$</td>
<td>number of documents in a corpus.</td>
</tr>
<tr>
<td>$O$</td>
<td>number of locations in a corpus.</td>
</tr>
<tr>
<td>$K$</td>
<td>number of topics (constant across time slices).</td>
</tr>
<tr>
<td>$L$</td>
<td>list of locations in a document (observed).</td>
</tr>
<tr>
<td>$l$</td>
<td>location assignment for topic $j$.</td>
</tr>
<tr>
<td>$z$</td>
<td>topic assignment for word $i$.</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>distribution of locations over topics.</td>
</tr>
<tr>
<td>$\phi$</td>
<td>distribution of topics over the terms.</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Dirchlet prior (hyperparameter) for $\phi$.</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Dirchlet prior (hyperparameter) for $\lambda$.</td>
</tr>
<tr>
<td>$w$</td>
<td>word (observed).</td>
</tr>
<tr>
<td>$t$</td>
<td>time.</td>
</tr>
<tr>
<td>$T$</td>
<td>length of time represented by the model.</td>
</tr>
</tbody>
</table>

Figure 6.1: Graphical model representation of the DSTM for three consecutive time slices.
2. Randomly draw $O_t$ multinomial distributions from $\lambda_t$, where $\lambda_{t,k} | \lambda_{t-1} \sim Dir(\lambda_{t-1}) + N(\mu, \delta^2)$.

3. For each document $d$, then for each word $w$ in the document:

   (a) Draw location $l$ and $z$.

   (b) Draw word $w$ from topic $z$.

Here $\phi_t, \lambda_t, z_t, w_t$, and $L_t$ are hidden variables, and $w_t$ and $L_t$ are the only observed variables. $\beta$ and $\delta$ are considered fixed here as recommended in literature for simplicity. The generative process for DSTM yields the distribution $p(\phi_t, \lambda_t, z_t, w_t, l_t | L_t, \beta, \delta)$ which can be decomposed according to the chain rule as follows:

$$p(\phi_t, \lambda_t, z_t, w_t, l_t | L_t, \beta, \delta) = \prod_{i=1}^{N_t} p(z_t, l_t | \lambda_t) p(w_t, l_t | z_t, \phi_t) \prod_{j=1}^{K} p(\phi_t | \beta) \prod_{y=1}^{O_t} p(\lambda_t | \delta)$$ (6.1)

The main inferential problem we are trying to solve is computing the posterior distribution of the hidden variables. To derive their posterior distribution from the joint distribution (Eqn. 6.1) we use Bayes’ rule:

$$p(\phi_t, \lambda_t, z_t, l_t, w_t | L_t, \beta, \delta) = \frac{p(\phi_t, \lambda_t, z_t, w_t, l_t | L_t, \beta, \delta)}{p(w_t, l_t | L_t, \beta, \delta)}$$ (6.2)

This distribution is very hard to calculate. To solve this problem we use an approximation technique.

### 6.2 Parameter Approximation

Given that we need to sample multiple parameters at once, Gibbs Sampling is an appropriate choice for approximating the hidden parameters. We can infer $\phi_t$ and $\lambda_t$ by first inferring the topic and location assignment $(z_t, l_t)$ pairs per word conditioned on all other variables
By applying Bayes rule we can obtain \( p(z_t, l_t) \) assignments as follows:

\[
p(z_t, l_t | w_t, z_{t,-i}, l_{t,-i}, w_{t,-i}, L_t, \beta, \delta) = \frac{p(z_t, l_t, w_t | \beta, \delta, L_t)}{p(z_{t,-i}, l_{t,-i}, w_{t,-i} | \beta, \delta, L_t)}
\]

This yields to:

\[
p(z_t, l_t, w_t | \beta, \delta, L_t) \propto p(z_t, l_t, w_t | \beta, \delta, L_t)
\] (6.3)

In the above equations, \( z_{t,-i} \) denotes the assignment of all topics except the current instance and \( l_{t,-i} \) denotes the assignment of all locations except the current instance. Now we can obtain the conditional probability of \((z_{t,i}, l_{t,i})\) based on \( z_{t,-i}, l_{t,-i} \) and \( w_t \) from equation 6.3 by integrating over the continuos distributions (Dirichlet distributions) \( \phi_t \) and \( \lambda_t \).

\[
p(z_t, l_t, w_t | \beta, \delta, L_t) = \int_{\phi_t} \int_{\lambda_t} p(z_t, l_t, w_t, \phi_t, \lambda_t | \beta, \delta, L_t) d\phi_t d\lambda_t
\] (6.4)

We can expand 6.4 using the joint probability distribution (Eqn. 6.1) and grouping terms by their dependent variables:

\[
p(z, l, w | \beta, \delta, L) = \int_{\phi} p(w | z, \phi) p(\phi | \beta) d\phi \int_{\lambda} p(z | \lambda, l) p(l | L) p(\lambda | \delta) d\lambda
\]

These two integrations represent a multinomial distribution multiplied by a Dirichlet prior.

\[
p(z, l, w | \beta, \delta, L) = \int_{\phi} \left( \prod_{i=1}^{N} p(w_i | \phi_{z_i}) \right) p(\phi | \beta) d\phi \int_{\lambda} \left( \prod_{i=1}^{N} p(z_i | \lambda_{y_i}) \right) \left( \prod_{m=1}^{D} \prod_{i=1}^{N} p(l_i | L_m) \right) p(\lambda | \delta) d\lambda
\] (6.5)
Working with the first term in Eqn. 6.5

\[ \int_{\phi} \left( \prod_{i=1}^{N} p(w_i|\phi, \text{j}) \right) p(\phi|\beta) d\phi = \]

\[ \int_{\phi} \left( \prod_{j=1}^{K} \prod_{n=1}^{V} \phi_{n,j}^{C_{n,j}^{\text{VK}}} \left( \prod_{i=1}^{K} \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \prod_{n=1}^{V} \phi_{n,j}^{\beta-1} \right) \right) d\phi \]

\[ = \int_{\phi} \left( \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \right)^K \left( \prod_{i=1}^{K} \prod_{n=1}^{V} \phi_{n,j}^{C_{n,j}^{\text{VK}}+\beta-1} \right) d\phi \]

\[ = \text{CONST}_1 \int_{\phi} \left( \prod_{i=1}^{K} \prod_{n=1}^{V} \phi_{n,j}^{C_{n,j}^{\text{VK}}+\beta-1} d\phi \right) \]

where \( \text{CONST}_1 = \left( \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \right)^K \) and \( C_{n,j}^{\text{VK}} \) is the number of times word \( n \) in Vocabulary \( V \) was assigned to topic \( j \) (among \( K \) topics).

\[ = \text{CONST}_1 \prod_{i=1}^{K} \int_{\phi} \left( \prod_{n=1}^{V} \phi_{n,j}^{C_{n,j}^{\text{VK}}+\beta-1} \right) d\phi \]

Given that the term \( \phi \) is a Dirichlet distribution, the resulting integral will be as follows:

\[ = \text{CONST}_1 \prod_{i=1}^{K} \frac{\prod_{n=1}^{V} \Gamma(C_{n,j}^{\text{VK}}+\beta)}{\Gamma(\sum_{n=1}^{V} C_{n,j}^{\text{VK}} + V\beta)} \]

(6.6)

where the Dirichlet integrals are obtained by applying the following rule:

\[ \int \prod_{x=1}^{X} \frac{a_{x}^{k_{x}-1}}{\Gamma(k_{x})} d\alpha_{x} = \frac{\prod_{x=1}^{X} \Gamma(k_{x})}{\sum_{x=1}^{X} k_{x}} \]

We can use the same machinery for the second term in Eqn. 6.5:
\[
\int_{\lambda} \left( \prod_{i=1}^{N} p(z_i | \lambda_{yi}) \right) \left( \prod_{m=1}^{D} \prod_{i=1}^{N} p(l_i | L_m) \right) p(\lambda | \delta) d\lambda \\
= \int_{\lambda} \left( \prod_{i=1}^{N} \frac{\lambda_{yi}}{1} \right) \left( \prod_{y=1}^{O} \left( \frac{\Gamma(K \delta)}{\Gamma(\delta)^K} \prod_{j=1}^{K} \lambda_{yj}^{\delta-1} \right) \right) d\lambda \\
= \int_{\lambda} \left( \frac{\Gamma(K \delta)}{\Gamma(\delta)^O} \right)^{O} \left( \prod_{m=1}^{D} \frac{1}{L_m} \right) \left( \prod_{y=1}^{O} \prod_{j=1}^{K} \lambda_{yj}^{C_{yj}^{KO}+\delta-1} \right) d\lambda \\
= CONST_2 \int_{\lambda} \left( \prod_{y=1}^{O} \prod_{j=1}^{K} \lambda_{yj}^{C_{yj}^{KO}+\delta-1} \right) d\lambda \\
\]

where \( CONST_2 = \left( \frac{\Gamma(K \delta)}{\Gamma(\delta)^O} \right)^{O} \left( \prod_{m=1}^{D} \frac{1}{L_m} \right) \) and \( C_{yj}^{KO} \) is the number of times location \( y \) was assigned to topic \( j \) (among \( O \) locations and \( K \) topics).

\[
= CONST_2 \prod_{y=1}^{O} \int_{\lambda} \left( \prod_{j=1}^{K} \lambda_{yj}^{C_{yj}^{KO}+\delta-1} \right) d\lambda \\
= CONST_2 \prod_{y=1}^{O} \frac{\prod_{i=1}^{K} \Gamma(C_{yj}^{KO} + \delta)}{\Gamma(\sum_{j'} C_{j'y}^{KO} + K \delta)} \\
\]

Substituting equations 6.6 and 6.7 in equation 6.5 we can obtain the following equation for \( p(w, z, l | \beta, \delta, L) \):

\[
p(z, l, w | \beta, \delta, L) = \\
CONST \left( \prod_{i=1}^{K} \frac{\Gamma(C_{nj}^{VK} + \beta)}{\Gamma(\sum_{j'} C_{nj'}^{VK} + V \beta)} \right) \left( \prod_{y=1}^{O} \frac{\Gamma(C_{yj}^{KO} + \delta)}{\Gamma(\sum_{j'} C_{j'y}^{KO} + K \delta)} \right) \\
\]

where
\[ \text{CONST} = \left( \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \right)^K \left( \frac{\Gamma(K\delta)}{\Gamma(\delta)^K} \right)^O \left( \prod_{m=1}^{D} \frac{1}{L^N_m} \right) \]

Finally, by substituting equation 6.8 in 6.2 and using the identity \( \Gamma(K+1) = K\Gamma(K) \), we obtain the Gibbs sampling

\[
p(z_{t,i}, l_{t,i}, w_{t,i}, z_{t,-i}, l_{t,-i}, w_{t,-i}, \beta, \delta, L_t) \propto \frac{C_{t,i,j}^{V,t} + \beta}{\sum_{i'} C_{t,i',j}^{V,t}} + \frac{C_{t,y,j}^{O,t} + \delta}{\sum_{j'} C_{t,y,j'}^{O,t} + K\delta}
\]

\( C_{t,i,j}^{V,t} \) is the count of word \( i \) from vocabulary \( V \) assignments to topic \( j \) from topics \( K \). \( C_{t,y,j}^{O,t} \) is the count of location \( y \) from locations \( O \) assignment to topic \( j \) from topics \( K \). The apostrophe on \( j \) and \( i \) denote all instances except the current one and \( t \) denotes the counts at time slice \( t \).

### 6.3 Model Evaluation

In order to quantitatively evaluate our proposed model, we compare its predictive power against other models. Our model inherits its dynamic nature from Dynamic Topic Model (DTM) proposed by Blei in [Blei and Lafferty, 2006]. DTM does not take into account any extra information in the process of inferring topics, and this gives our model an advantage over it. Another parent to our model is the Author Topic Model (ATM), a non dynamic model proposed by Rosen in [Rosen-Zvi et al., 2004]. Our model inherent the ability of integrating extra information into the topic inference from ATM. In our case, the extra information is the location. ATM is non dynamic, which also gives our model an advantage over it.

We compare our model against ATM and our baseline model is LDA. For the purpose of comparing the three models, we ran each model separately on the same dataset (described in 2.2) while fixing the hyper parameters of the model. The hyper parameters are fixed as follow: \( \beta \) is calculated as \( 50/K \) and \( \delta \) is fixed to 0.01, where \( K \) is the number of topics. The comparison was based on calculating the perplexity for each model while varying some model parameters. The
perplexity score for an unseen document conditioned on observed locations is calculated as follow:

$$\text{Perplexity}(w|l) = \exp\left(-\frac{\log p(w|l)}{N}\right)$$  \hspace{1cm} (6.9)

where $p(w|l)$ is the probability that the word $w$ appear in the unseen document conditioned on the location observed from the document and the pre-trained model. $N$ is the total number of words in the unseen document. To calculate the perplexity of the testing set, we average over the documents as follow:

$$\text{Perplexity}(D_{\text{test}}) = \frac{\sum_{d=1}^{D} \text{Perplexity}(w|l)}{D}$$  \hspace{1cm} (6.10)

Where $D$ is the number of documents in the test set.

Comparing our model against LDA model with respect to the vocabulary size (number of unique words in dataset) revealed improvement in perplexity to our model advantage when the dataset size grows bigger. Fig.6.3 reports on perplexity with respect of vocabulary size. Both models show slightly close performance until the size of the dataset gets bigger.

Measuring performance while varying the number of topics our model shows better performance than LDA model with fewer number of topics Fig.6.2. The performance of our model decreases with a higher number of topics. The results show that the optimal number of topics is under 10 topics.

### 6.4 Model Applications

Here we take a qualitative approach to evaluating our model by exploring the applicability of our model on historic newspapers. In this section, we present two applications. The first application is concerned with news coverage of three newspapers from the east, midwest, and west. The model output was used to understand the differences in reporting between these newspapers. The
second application focused on understanding the tone usage in relationship to discovered topics
and locations. In this application, the dataset was divided based on tone resulting in four different datasets. Each dataset was fitted to the model and the output was analyzed and compared.

6.4.1 East, west, midwest 1918-1919 news coverage

In this application, we ran the model on three individual historic newspapers. We chose the newspapers randomly from three different areas of the United States: east, mid-west, and west. Newspapers used in this study were: New York Tribune, NY from east, The Evening Missourian, MO from mid-west, and Bisbee Daily Review, AZ from west. We analyzed the topics and locations discovered from these newspapers and then mapped to historical facts reported in published reports by the Navy Department Library [Rep, c], United States Department of Health and Human Services [Rep, b], and National Archives and Records Administration [Rep, a]. We extracted influenza paragraphs from the three newspapers and then divided them into months. In this application, we discarded European locations in the location detection phase because the main focus of this research is the United States influenza pandemic. We also discarded paragraphs with no location mentioning and focused on explicitly mentioned locations. We visualized topics and locations from September 1918 through January 1919 for the output analysis. To highlight locations from the three different US areas, we color coded locations as following: Green to denotes cities in and around the west, Black to denotes cities from mid-west, and Blue to denotes cities from east.

Examining the topics from New York Tribune September 1918 segment the first topic/locations indicate that cases of pneumonia were reported by the health department in Manhattan, NY and Somerville, NJ. Officials at Boston and Washington reported ill people with influenza and death cases. They also reported on their concerns of the spread of Spanish influenza disease. There were reports on the spread of Spanish influenza in the German army, and then to England. Reports from New York, NY on establishing a quarantine to stop the disease spread. The second topic in the same segment included the words Copeland, commission, steamer, symptom, nasal, germ, and isolate because Copeland, city health commissioner, reported on the isolation of people who came to the states (from France) by a steamer. Copeland also said symptoms of Spanish influenza include nasal
discharges and that the germ is carried in the nose and mouth. Again here the appearance of the locations Minnesota, Iowa, California, Connecticut, and Camp Lee, VA was due to reports on the number of cases, concerns and deaths.

In the January 1919 segment first topic, the words pathetic and wald emerged because there were some press releases by miss Wald, founder of Henry Street Settlement in New York, on the difficulties faced because of the influenza situation. In the second topic, some words related to advertisement showed up in this topic as color, main, and floor. Example on an advertisement: “Snug Bath Robes - Such a comfort to lounge in, such protection, too, when severe cold seems such an epidemic. $4.89 to $19.74 Macy’s Main Floor, 35th Street.” The words authority, roosevelt, association, and loss, mead are due to reports from authorities from Bulgaria on losses due to influenza around the same time Franklin d. Roosevelt, assistant secretary of the navy, contracted influenza on his trip to Europe. In January 1919, there were reports on the death of Theodore Roosevelt who served as the 26th President of the United States. S.C. Mead, the secretary of the merchants’ association, commented on his death.

From The Evening Missourian newspaper results we examined closely on the following months: September, October, and December of 1918. During the month of September 1918, the topics and assigned top locations were the results of major reporting on a liberty loan parade happening in Columbia, MO. Schools were kept out of the parade to minimize the spread of the disease. Around the same month Ferguson tablets were advised to be taken for cold and grippe. The President of Columbia board of health announced news on spanish influenza cases. There were reports on an influenza outbreak and death cases reporting in the great lakes naval training school and station. The word enemy and location New York emerge in the second topic due to reports on an attempt of bombing New York by the enemy that was Germany at this time. Since this topic had some war related words this explains the emerging of the location Springfield, MA, were an army camp was located, in the topic top five locations. The rest of the locations had reported influenza cases or related deaths.

In October’s topics and locations. Influenza cases were reported in New York, West Virginia,
and at the parker memorial hospital. Troops were supposed to go to Camp Bowie in October, but they did not because of the epidemic. Board of health in San Francisco, California demanded people to wear masks. In the second topic the words nurse, school, and ill appeared as an indication of reports on influenza cases in nurses school. The words student and house appeared in this topic because students were advised to stay in their houses and not go to school. Also, there were reports on students of the school of medicine nursing influenza cases at hospitals. Announcements from the army were made through the Arlington naval radio station. This explains the emerging of the location Arlington in the second topic. Many reports on influenza cases in the army and some reports on the secure arrival to the US from sea.

In December 1918 topics, topic one contains the words red, cross, receive, Christmas, service, influenza, soldiers, and epidemic. Those words indicate the involvement of the red cross in helping the soldiers during Christmas time. In the second topic, Kansas city rises in the top five locations. The second topic contains the words vaccine, university, louis, student, and die because there were some death cases. These locations and set of words confirms with the reports, during December 1918, on the availability of the vaccine in Kansas city and St. Louis. The university of Missouri developed the vaccine and served as the main source of the vaccine. There were demands to release doctors from army in Illinois to match the needs of the community. Des Moines, Spokane, Philadelphia also appeared in the top locations due to reports on ill and dying people with influenza.

In the Bisbee Daily Review Newspaper, the September 1918 first topic terms and locations confirmed that the epidemic of spanish influenza and pneumonia appeared in army camps, and it is causing many deaths. As a result, Camp Meade was placed under general quarantine. Reports from mass. (Massachusetts) on the outbreak and number of new cases reported from Camp Devens, MA. At this time, Massachusetts was believed to be the center of the epidemic in the east. Reports from Richmond, VA health officers on their measures to prevent the spread of spanish influenza. There were also reports on the state of the influenza in Connecticut, CT, and New Hampshire, NH. From the second topic/locations, the terms were a guidance to many facts. In Norfolk, VA cases of spanish influenza have developed among enlisted men at the Hampton roads naval base. Influenza spread over the country and reached Boston in September. In this month, several cities and towns within a
twenty five mile radius of Boston reported deaths from influenza and pneumonia. The commissioner of Boston gave some press releases to tame the panic of the public. One of his statements was: “fear would lower the vitality of those exposed”. On the other hand, cases of epidemic reported in army technical school at the university of Colorado. It was reported by review leased wire in Washington that El paso, phoenix, and jerome closed all public places. There were also reports on deaths in Montana and Warren district. The words enemy, french, british, attach, and german emerged in the same topic which confirm major war related reporting during this month. Reporting on American troops attacking the west of the Verdun region in co-operation with the french. Around the same time, British troops made a powerful attack against the german lines.

In January 1919, the second peak of influenza was striking. The words in the first topic are related to some news about Campbell, Arizona’s governor. His sons were seriously ill, and he was leaving the office. The word attack appeared because spinach influenza was described as attacking people and resulted in their death. There was many reporting on the death in New England. Number of locations like Seattle, Arkansas, and Maricopa were assigned to this topic because during this time Seattle Washington former mayor died from influenza. Chandler William Clemens who was born in Arkansas also died after an attack of influenza. There were reports on officials visiting the county of Maricopa and the influenza epidemic situation. During this period, it was believed and announced by authorities that spanish influenza is just grip camouflaged under a new name which explain the appearance of the word grip in the same topic. Again in the second topic it is very obvious that influenza attacks are happening. There were reporting on the attacks in Douglas, a city in Arizona. Reports during this month was as follow: “a fresh outbreak of influenza is shown in phoenix by the report of 117 new cases during the last 72 hours.” Placards were being posted on all the houses in the warren district where cases of influenza were found. There was some weather related reporting from Holbrook during this period. The words war and court appeared in this topic due to reporting on collecting of war profits and the involvement of the supreme court in the process. The supreme court was also mentioned in freedom of speech related reports. Both supreme court reports were not directly related to influenza but happened around the same time.

A great number of reports on the 1918-1919 influenza epidemic were published by the Navy
For clarity of reporting and the specific mentioning of areas where the pandemic appeared we closely followed the report published by the Navy Department Library under the title "The pandemic of influenza in 1918-1919". The report mentioned two major statements:

1. First statement: “The peak of the epidemic was reached in September in Navy personnel and about the middle of October in the Army.”

2. Second statement: “In September, it appeared in rapid succession in other Army camps and the civilian population along the Atlantic seaboard and the Gulf of Mexico and spread rapidly westward over the country.”

These two statements mention the peak of the pandemic and where it started and to which directions it did spread. We examined the topics and locations starting September 1918 through January 1919 in the three newspapers shown in Fig. 6.4, 6.5, and 6.6.

In the three newspapers September and October topics navy and war related terms like war, army, attack, and naval emerged in top 15 terms. Other epidemic related terms like report, death, pneumonia, spread, and disease also emerged. The two sets of terms emerging in September and October time slices confirms with the first statement in the report.

Examining the top 5 locations in each time slice from the three newspapers location camps bubbled up to the top locations. Also, locations from the midwest and west of the country started appearing in later time slices that confirm with the second statement. Examining the output, we found that all three newspapers mainly reported on local and surrounding areas. Few national (any
location outside the newspaper publishing state) locations have emerged in the top five locations. The emerged national locations in the three newspapers do confirm with the disease spread over the country. Locations from the east emerged in the September and October time slices in the midwest and west newspapers. On the other hand, in the east newspaper there were no national locations, and this indicates the intensity of reporting given that the epidemic at this time was concentrating in this area. Similarly, locations from the west emerged in January 1919 in the east and midwest newspapers, and this confirms with the epidemic spreading toward the west during this time.

The analysis of the topics and locations discovered from New York Tribune, NY from east, The Evening Missourian, MO from mid-west, and Bisbee Daily Review, AZ from west confirmed with major events in reports on the epidemic. Mapping the timeline of influenza peaks and spread was also confirmed.

6.4.2 1918-1919 Influenza related tones, topics, and locations

In this application, our main goal is to understand the shifts of topics based on tones used by newspapers from different parts of United States. Newspapers here were also divided into three
We present a supervised learning approach to tone detection. Finally, we explain the added insight offered by detecting the tone prior to applying our model.

Tones used in this application were identified by four domain experts: one historian, one librarian, and two rhetoricians. We focused on four main tones: alarmist, explanatory, reassuring, and optimistic.
and warning. They first identified tones from advertisements then applied them to a broader sample of texts. Here is a brief description of tones:

1. Alarmist: uses fear or urgency, often mentioning number of sick or dead; induces a sense of panic.

2. Reassuring: comforting; implies threat is diminishing; addresses fears with soothing sensibility; typically conveys the idea that if one takes a recommended action; motivates action with a sense of hopefulness, improvement, or possibility of avoidance of disease; involves sense that an action will lead to betterment.

3. Explanatory: discourse as a source of information; lacks a distinctive affect.

4. Warning: serious but not urgent; cautioning; advises the reader what to do; mentions measures being taken, but conveys no sense that the threat is diminishing.

For the classifier, we used a Multinomial Naive Bayes classifier. It is first trained using the features extracted from label data. For features extraction we used a Tfidf from a 1- to 2-grams language model. Using the labeled data, we trained the classifier to detect the four tones (alarmist, explanatory, reassuring, and warning), using approximately 300 cleaned sentences from newspapers and four coders, who attained a moderate level of agreement in their classifications (Kappa=0.47). Kappa is a measure used to assess the degree of agreement among raters.

The dataset used here is the same data described in the previous section. After running the tone classifier, we use the results to divided the data into four datasets, one for each tone. We then apply our model on each dataset separately and compare the results. Figure 6.7 shows the different tones distribution over Influenza reporting starting January 1918 until December 1919. The Explanatory tone peaks when the Influenza peaks. The gap between the explanatory tones peaks is bigger than other tones in the west and midwest, but not in the east. The east coast reporting seems to exhibit a variety of tones, but the explanatory tone still dominant. The alarmist tone is the least prominent tone in the three parts. The explanatory tone is the most prominent tone in west,
midwest, and west. This is because the dataset we are applying the tone classifier on is a newspaper dataset and newspapers tend to be neutral in their reporting.

Applying the DSTM model resulted in topics and locations for each time slice in each tone. Here we show the output for east coast newspapers. Figure 6.8 is a visualization of the model output for each tone starting September 1918 until January 1919.

![Figure 6.7: Tones distribution over Influenza reporting.](image)

(a) West coast.

(b) Midwest.

(c) East coast.
Figure 6.8: East coast newspapers discovered topics and locations grouped by tones.

As expected for east coast newspapers, the east locations dominated the maps in every month for every tone. The interesting pattern to notice is the different locations appearing in different tones and times. More locations appeared from the west and midwest in the explanatory and reassuring
tones. This is because east coast newspapers still report about the situation in west and midwest, but in an explanatory and/or reassuring tone. Given that the epidemic reached the west around January 1919, west coast locations started to appear in alarmist, warning, and reassuring, but the situation did not change for explanatory.

In the alarming output, words suffering, rage, and thousands appeared in topics. These words did not appear in any of the other tones. In the month of September 1918 the first peak of the Influenza epidemic was around this time which explain the appearance of the words epidemic, influenza, death, and increase. At this time, influenza was described as "the dim and ineffectual wraith of the mediaeval black death". It was reported by all authorities that 1510 bubonic plague (historically called black death) was the influenza. There were also arguments that the plague that devastated the Greek armies at the siege of troy was influenza which explains the grouping of the words disease, wraith, mediaeval, ineffectual, and death in one topic. Top locations were only in the east, and that was expected because it was the first peak of the epidemic and it started from the east side of the country. In December 1918, topics were result of reports of an influenza outbreak in some west coast locations and medical supplies were shipped to them. This explains the emerging of west locations in the top locations. Around the same time, there were reports on the arrest of a post office employee who was extracted money from envelopes. December is Christmas time in which families send money to each other as a gift. The man was caught by post office Inspectors.

Examining the reassuring output we found words like preventive, laxative, tablet, sore, nausea, tea, germ, and cough. Those set of words show that there were a great number of reporting on Influenza and grip symptoms and some preventive measures. For example in November 1918 there were reports on laxative bromo quinine tablets that were advised to be taken as a preventive measure from grip. Liberty catarrhal cream was also advised to be used to kill germs in the nose throat and intestines also as a preventive of influenza and other infectious diseases. People were advised to never let a cough or cold or tease of grippe get serious. A spray called “Tonsiline” was also advised for gargle to relieve sore throat upon its first appearance. None of the other tones has this amount of preventive measures related words. The goal behind all these preventive measures were to reassure the public that there were hope that they can escape the disease or at least easy its
symptoms.

From the warning output, the word spread appeared in most time slices. For example, in September 1918 words like influenza, speed, spinach, and serious appeared together in one topic which indicates the size of the problem. As a result, there were reports referencing doctors as: Dr. Phelps and Dr. Fowler to warn people from being in crowded places. They advised people to say at home. There were also reports on a grip available medicine ordered by doctors and that it does not cure it. The word drink emerged in the same topic the word flower and patient because there were reports from Dr. Flower on where hot drinks are good for influenza patients or not.

During January 1919, there were reports in the form of warnings. The words weak, cough, and vitality appeared in one of the topics as a result of many reporting on the influenza symptoms. There were statements in Washington on an available remedy it was called "hypo-cod". Again January was around the time when the epidemic reached the west.

Detecting tone and dividing the dataset based on it gave a different perspective to discovered topics and their locations. Topics discovered from the same time slice in one tone were different from others discovered from another tone. Adding the tone aspect gave us a chance to see the data from a different angle and derive different conclusions.

### 6.5 Summary

Here we introduced a probabilistic model that can model relationships between locations, topics, documents, and terms in a dynamic fashion. The model enables summarizing and navigating unstructured time stamped documents while capturing the evolution of topics along with location distribution over these topics.

We presented two different applications of the DSTM. The first application focused on understanding the differences in news coverage between the east, west, midwest parts of U.S. in 1918 and 1919. The second application focused on discovering the differences in news reporting
between the three parts of U.S. from the reporting tone perspective.

We evaluated the DSTM qualitatively and quantitatively. The quantitative evaluation was done by comparing our model to existing models using perplexity. The model showed better performance over the basic topic model (LDA) and slightly better performance than the Author-Topic model (ATM). We qualitatively evaluated the model by closely examining the output from the two applications, described above, and mapping results to published reports on the influenza epidemic peaks and spread pattern. Overall the model was successful in identifying the trends and their locations which helped in studying the difference between news reporting in the two applications.
Chapter 7

Predictive Analysis

In this chapter, we will present the fourth and final part of this dissertation, which is enabling the Dynamic Spatial Topic Models (DSTM) for predictive analysis. The motivation behind this work is to describe a powerful model that predict a major event and where this event will happen from unseen documents. Documents can be in the form of social messages, newspaper articles, blogs, or any form of textual data. The main research question we are trying to answer is: how can we predict what and where a major event will happen? To predict future topics and their locations from unseen tweets, we have adapted the work from [Wang et al., 2012], in which they proposed training a basic topic model (LDA) using data from seven days to calculate a transition parameter form discovered topics. This transition parameter is then used to predict the distribution of topics in unseen tweets from the 8th day. The transition parameter needs to be updated every time new data is streamed.

There are two major drawbacks of the Wang et al. work: it is based on basic LDA (a non-dynamic and non-spatial topic model), and updating the transition parameter is computationally intensive. In this part of our work, we overcome those drawbacks by training the model using DSTM (our model) instead of LDA. Using DSTM, enabled topic and location discovery from data collections and there is no need to update the transition parameter, since our model is dynamic; topics discovered at time $t$ are evolved from topics discovered at $t - 1$. Although the resulting framework is broadly applicable, we apply it primarily over our Latin American tweet collection,
previously described in 2.4.

### 7.1 Prediction Approach

The prediction approach we are presenting here is for streaming data. We train the model from seven days by applying DSTM on the available data then calculate the transition parameter. Using the transition parameter, we can predict topics from unseen tweets from the 8th day. Figure 7.1 shows three sample runs, showing the data that will be included in the transition parameter calculation phase and data used in the prediction phase for each run. In Wang approach, in the second run, the transition parameter will be updated using a computationally intensive method, discussed in [Wang et al., 2012].

Each run consists of a document collection $D$ of seven days worth of data. $D$ consists of $m+1$ documents. We apply DSTM on the documents to generate the topic-term distribution and the location-topic distribution. Here a document refers to a location-topic distribution, where each row represents a topic distribution over the locations.

The training data $D$, used to calculate the transition parameter, is divided into $D_{old}$ and $D_{new}$. $D_{old}$ is the collection of documents that will be used for training and $D_{new}$ is the collection of documents that will be used for testing. In more details:

\[
D_{old} = D_t(1,m) \quad (7.1)
\]
\[
D_{new} = D_t(2,m+1) \quad (7.2)
\]

As in any prediction problem we need to define/calculate the prediction error and attempt to minimize it. Some researchers use iterative methods (e.g. gradient decent), but we use an optimal solution to minimize the prediction error is using a direct method, previously presented at [Wang et al., 2012]. We calculate the prediction error, as follows:
For this run the topics discovered from the 8th day evolve from topics discovered from 7th day from Run 2

For this run the topics discovered from the 8th day evolve from topics discovered from 7th day from Run 1

$\text{Figure 7.1: Experimental setup for predicting topics and their locations from streaming data.}$

$$\text{error}_{\text{prediction}} = \min ||D_{\text{predicted}} - D_{\text{actual}}||_F^2$$

where $|| \cdot ||_F^2$ is the Frobenius matrix norm, $D_{\text{predicted}}$ is the predicted location-topics distributions, and $D_{\text{actual}}$ represent the actual location-topic distribution. The predicted location-topic distributions is calculated as follows:

$$D_{\text{predicted}} = D_{\text{old}} \times TP$$

where $D_{\text{old}}$ represent the old location-topic distribution, s.t. each row in this matrix represent a
location distribution over topics. TP is the transition parameter.

Transition parameter TP, is a matrix of size $K \times K$ where $K$ is the number of topics. Number of topics in $D_{old}$ and $D_{new}$ should be the same and should vary based on the application. We calculate the transition parameter as follows:

$$TP = (D_{old}^\prime D_{old})^{-1}D_{old}^\prime D_{new}$$

Here $D_{old}^\prime$ represent the transpose of $D_{old}$, $(.)^{-1}$ represent a Cholesky Factorization. Unlike the transition parameter calculation approach presented in [Wang et al., 2012], here we don’t need to update the transition parameter because DSTM is a dynamic topic model. The same TP equation will be used for every 7 days worth of data when they become available.

To predict the topic distribution over the unseen documents we use documents from the 8th day. Documents here are actual documents unlike the analogy we used previously. We divided the 8th day document set into previous documents and future (unseen) documents sets. The prediction process steps are:

- For each document in the previous document dataset:
  - we infer the document-topic distribution using the DSTM.
  - Then we multiply the resulting location-topic distribution from the previous tweet by the transition parameter to get the predicted topic distribution of the future (unseen) tweet.

In the inference phase, our main goal is to calculate the conditional probability of a document given a topic $p(topic|document)$. We can calculate this probability by inferring the document-topic assignment as follows:

$$p(z|d,l) = \sum_w p(z|w,l)p(w|d)$$

where $d$ represents a document, $w$ represents a word in a document, and $z$ represents a topic. $p(w|d)$
is the normalized word \( w \) frequency within document \( d \). We can calculate \( p(z|w) \) given the trained model as follows:

\[
p(z|w, l) = \phi_{wz} \ast \theta_{lz}
\]

Where \( \phi_{wz} \) is the word \( w \) probability to appear in topic \( z \) retrieved from the topic-terms distribution \( \phi \), and \( \theta_{lz} \) is the topic \( z \) probability to appear in location \( l \) retrieved from location-topic distribution \( \theta \).

The result of the previous process is the predicted topic distribution for the unseen documents. Now we examine the predicted topic distribution to find a topic with the highest concentration and assign it to the predicted (unseen) document. Table 7.1 is a sample topic assignment for the unseen documents mapped to the actual documents. In this example, the document is a preprocessed tweet with a partial set of words shown for each tweet. These tweets were not observed at the prediction time, but they are shown here for illustrative purposes.

After having an assigned topic for each unseen document, we can then count each topic assignment and rank the topics from highest assigned to lowest assigned. Table 7.2 depicts an example of the topic assignment counts. From this table, topic one is the topic with the highest count, and this indicate that it is the most probable topic to appear in the unseen documents on June 8th, 2014. The terms and locations for a topic can be retrieved from the topic-document distribution and location-topic distribution discovered from the trained model.

In the following section we will discuss a use case of the proposed framework and a detailed discussion of the results will be presented as an illustration on how to interpret the output.

### 7.2 Latin America Unrest Prediction

In this application, the main goal is to predict civil unrest events happening in Latin America. The focus here was on Twitter data collected from: Colombia, Mexico, El Salvador, Costa Rica, Guatemala, Chile, Paraguay, Argentina, Venezuela, and Ecuador. This data was provided by the
Table 7.1: Sample topic-document(tweet) assignment

<table>
<thead>
<tr>
<th>Date</th>
<th>Country</th>
<th>Tweet Terms</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>6/8/13</td>
<td>Colombia</td>
<td>consigue - mundo - punto - empate - colombia - arquero - argentina</td>
<td>4</td>
</tr>
<tr>
<td>6/8/13</td>
<td>Argentina</td>
<td>música - escuchar - pide - loco - tomatela - celular</td>
<td>1</td>
</tr>
<tr>
<td>6/8/13</td>
<td>Argentina</td>
<td>salvo - juega - cancha - james - seguir - argentina</td>
<td>2</td>
</tr>
<tr>
<td>6/8/13</td>
<td>Colombia</td>
<td>objetivo - logro - tricolor - listo - vamoscolombia - sisepuede - fcfselec-</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cioncol - sumar</td>
<td></td>
</tr>
<tr>
<td>6/8/13</td>
<td>Venezuela</td>
<td>leña- olor - quedas - aquel - campo</td>
<td>2</td>
</tr>
<tr>
<td>6/8/13</td>
<td>Chile</td>
<td>igualar - clasificación - colombia - posterga - mundial - argentina</td>
<td>1</td>
</tr>
<tr>
<td>6/8/13</td>
<td>Ecuador</td>
<td>messibelievers - ecuargentinaños - quieren - matar</td>
<td>2</td>
</tr>
<tr>
<td>6/8/13</td>
<td>Colombia</td>
<td>mejores - ganamos - perdimos - colombia</td>
<td>4</td>
</tr>
<tr>
<td>6/8/13</td>
<td>Chile</td>
<td>wazees - muuuuuucho - acabo - reportar - gauss - miguel - vehículo -</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>social- san - congestionamiento - gps</td>
<td></td>
</tr>
<tr>
<td>6/8/13</td>
<td>Ecuador</td>
<td>eliminatorias - partido - argentina - termina - colombia</td>
<td>2</td>
</tr>
<tr>
<td>6/8/13</td>
<td>Colombia</td>
<td>hpta - asistencias - pasayo - colombia - messi - pasao - quie - enano -</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>triple - juego - ajajajajaja - monda - malparido - goles</td>
<td></td>
</tr>
</tbody>
</table>

Early Model-Based Event Recognition with Surrogates (EMBERS) a Virginia Tech based project. The dataset used here were first introduced in 2.4. We filter tweets from these countries using a civil unrest related keywords.

In this application, location is assumed to be the location the tweet was initiated from. As a result, there is only one location assigned to each tweet. Number of topics was fixed to five. Results from our framework are represented as the top two topics assigned to the unseen tweets. Reference events for specific dates are retrieved from the reference event dataset previously discussed in 2.4. Reference actual unrest events and their locations were also provided by Early Model-Based Event Recognition with Surrogates (EMBERS). The top two topics are manually examined to validate their similarities with the reference events dataset.

Two examples on the framework output are explored as an illustration on how we interpret the output. The first example is focused on the first week of June, 2013. The first seven days are used for calculating the transition parameters, and the prediction was done for the unseen tweets from June 8th, 2013. Table 7.2 shows the resulting counts of the topics assignment to the unseen tweets. From this table the top assigned topics are topic one then topic four, Figure 7.2 is showing the top terms (translated to english) and top locations for each topic. Examining the top terms from
Table 7.2: Predicated topic assignment counts for June 8th, 2013.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 0</td>
<td>4,126</td>
</tr>
<tr>
<td>Topic 1</td>
<td>24,932</td>
</tr>
<tr>
<td>Topic 2</td>
<td>14,888</td>
</tr>
<tr>
<td>Topic 3</td>
<td>17,006</td>
</tr>
<tr>
<td>Topic 4</td>
<td>23,945</td>
</tr>
</tbody>
</table>

topic one and topic four and their locations, we found out that they confirm with the reference events (during the same day). Here are some notable observations:

Topic one includes the terms freedom, expression, people, and pathway. These terms are a clear indication on an unrest and a protest happening. Other terms such as water and environment were present along with the previous words which confirm with a reference event about group of people protesting the lack of water and environmental damage on June 8th, 2014. This event happened in Chile, which is one of the top locations of topic one. Work, money, and power also appeared in topic one and this also confirms with a protest related to street traders not allowed to earn a living. This protest event is one of the events in the reference events of the same day, and the actual event location was Mexico, also appeared in the topic top locations.

In topic four the terms hate, evil, and some inappropriate terms are an indication of public anger. After examining the reference events, we determined that during the same day there were a rally by the members of the council of the Sexual Diversity Mexico State and Lesbian Gay community seeking to legalize marriage between same-sex individuals and to criminalize hate crimes and homophobia. We posit that the anger related words might be a public reaction to this protest. These protests happened in Mexico, which is in the top locations of topic four. The term school appeared in the same topic, and this coincides with a protest event related to school teachers minimum wage. The protest happened in Venezuela and also Venezuela appeared in the top topic locations.

Another example is topics predicted from June 29th, 2014 unseen tweets. After predicting the topic assignment of the unseen tweets, counts can be calculated, and they are shown in table 7.3.
Figure 7.2: Predicted topics and their locations from the 8th day of June 2013

Topic three and topic two appear to be the most prominent. Figure 7.3 shows the top two topics assigned to unseen tweets from this day.

Figure 7.3: Predicted topics and their locations from June 29th, 2013.

Manually examining the topic terms and their locations along with the reference events from the same day we found the following: In topic three the terms school, poor, and wrong appeared together because at this day school teachers were protesting for living wage. The term school
Table 7.3: Predicated topic assignment counts for June 29th, 2013.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 0</td>
<td>15,365</td>
</tr>
<tr>
<td>Topic 1</td>
<td>13,060</td>
</tr>
<tr>
<td>Topic 2</td>
<td>18,830</td>
</tr>
<tr>
<td>Topic 3</td>
<td>18,919</td>
</tr>
<tr>
<td>Topic 4</td>
<td>16,731</td>
</tr>
</tbody>
</table>

appearance along with the words people and best also confirms with another march by high school students supporting the progress in education. There were also some protests by teachers about secondary education reform laws. This event happened in Mexico that appeared in topic three top locations. The terms sos, independent, and hatred are an indication to the feelings that surrounded these events.

In topic two the terms god, faith, savior, and glory is a good indication on a religious march reference event called “the March for Jesus” that happened during the fourth week of June. This event took place in Brazil, which is one of the top locations for topic two. Also, the appearance of the terms wrong and penalty in the same topic can indicate the feelings some people had toward the protests and marches by members of the gay, lesbian, bisexual, transgender and intersex (GLBTI). Both marches happened in the same day. One of these protests happened in Mexico (one of the topic two locations). Other protests happened in Ecuador and El Salvador, but these two did not appear in the top locations.

### 7.3 Summary

Here we presented a prediction approach for predicting topics from unseen documents. We applied the approach on Latin America tweets dataset with the goal of predicting civil unrest events and their locations. The applications severed as a qualitative evaluation for the prediction approach in which we compared the output to a reference event dataset. This event dataset consisted of actual unrest events and their reported time and location. Our prediction approach was successfully in predicting major events and their locations from unseen tweets.
Chapter 8

Conclusion

The main goal behind the work presented in this dissertation was improving the basic topic model (LDA) to help extract greater information from data and improve the utility of the text mining process. Each part of this work was motivated by answering a specific reach question. The research questions were the results of formalizing real problems faced by social scientists and humanists dealing with the ever-growing availability of digital archives and data from social media tools. To realize this goal we presented and evaluated the following:

1. A Dynamic Temporal Segmentations over Topic Models Algorithm, A time series segmentation algorithm that segment time based on shifts in topics.

2. ThemeDelta, a visual analytics system for discovering and representing the evolution of trend keywords into ever-changing topic aggregations over time.

3. Dynamic Spatial Topic Models (DSTM), a new model that incorporate reporting locations of inferred topics, and captures their evolutions over time.

4. Prediction approach, an approach for predicting topics and their location from unseen documents.
All four parts were successful in assisting humanists and social scientists in answering questions and drive conclusions. They offer great improvement specifically to the basic Topic Model (LDA) and to the text mining process generally. While these solutions are greatly successful, they also have imperfections and large space for improvements. In the following section, we will address each part in details and address its advantages, disadvantages, and suggest future modifications and extensions.

8.1 Dynamic Temporal Segmentations over Topic Models

This algorithm was successful in extracting important qualitative features in the form of topics and the duration in which these topics were present from a text corpus. The main idea behind this algorithm is seamlessly wrapping a time series segmentation algorithm around a topic molding algorithm. This flexibility allows the freedom of choosing the appropriate topic model to fit the problem at hand.

This algorithm has limitations stemming from non-adaptive window sizing and the fixed number of discovered topics. The minimum and maximum window sizes have to be pre-specified before the segmentation algorithm is run. This introduces the problem of force-adding a segmentation point when the maximum window size is reached, to overcome this we increase the maximum window size which can result in a slower running algorithm. The fixed number of topics discovered from each segment can introduce redundant topics.

Future work direction for this part can focus on extending the segmentation algorithm to capture not just topic differences but sentiment evolutions. For example, capturing the sentiment evolution will enable us to measure differences in public perception and attitudes between advantaged and disadvantaged neighborhoods. We can also address the pre-specified windows sizes and the fixed number of topics limitation as another direction of future work.
8.2 New Visual Analytics Representations

ThemeDelta excels at discovering trending topics and visualizing not just the discovered topics, but also their evolution over time. We have demonstrated the utility of our analytics component by applying it to several types of text corpora. However, while ThemeDelta has many strengths, it is also balanced by several weaknesses and areas of future improvement.

For the visualization component, limitations appear in the presence of many trends, long time periods, and high visual complexity. While existing techniques such as TextFlow [Cui et al., 2011] take a macroscopic approach to summarizing massive text corpora using high-level overviews, ThemeDelta uses a trend-level design that does not scale as well when the number of trends or time segments increases. While we have not derived a formal limit, even many of the examples presented in the New Visual Analytics chapter skirt the boundary of the utility of the technique. In practice, large datasets (in either trends or time, or both) yield high visual complexity, particularly in the number of trendline crossings as well as incident trendlines. Such effects make perceiving the visualized data more difficult. Several possible strategies can solve this problem, such as filtering, sampling, or aggregation.

We opted to not run a controlled quantitative experiment using ThemeDelta, opting instead for a qualitative expert review [Tory and Möller, 2005]. One reason for this choice is that we found no suitable technique to use as a baseline comparison for such an experiment. While techniques such as TextFlow [Cui et al., 2011] and TIARA [Wei et al., 2010] do provide insight on trends evolving over time, they cluster keywords together and focus on providing overview instead of detail at the level of individual keywords. In this sense, parallel tag clouds [Collins et al., 2009b] are perhaps the closest technique to ThemeDelta in that it visualizes individual keywords, yet PTCs do not show the clustering of trends into topics over time. This makes direct comparison difficult. While our qualitative review did not compare ThemeDelta to other techniques, it did give rise to much more qualitative and generally useful results.

Our future work will study aggregation methods—time-based and keyword-based alike—for
ThemeDelta that would increase the scalability of the system. Another focus will be to enable the automatic detection and visualization of the diffusion of ideas in scientific communities. From an algorithmic perspective, we will be looking to extend the segmentation algorithm ability to work with nonparametric bayesian models as Hierarchal Dirichlet Process (HDP) proposed by [Teh et al., 2006]. Unlike LDA, number of topics in the HDP model is automatically inferred. Currently, our algorithm only supports parametric Latent Dirichlet allocation given that they are the most commonly used in practice and research.

8.3 Dynamic Spatial Topic Model

DSTM succeeded in discovering the relationships between locations, topics, documents, and terms in a dynamic fashion. By applying the model on two text corpora, we demonstrated the model ability to summarize and navigate unstructured time stamped text documents while capturing the evolution of topics along with location distribution over these topics. We also quantitatively evaluated our model by comparing it to LDA and Author-Topic Model. One of the advantages of our model over these two models is that it companies the power of both. Another advantage is that our model does have better performance than LDA and a slightly better performance better than the Author-Topic model.

This model has two major limitations. First limitation is the fixed number of topics, also appeared in our segmentation algorithm. This limitation exists here because this model is an extension of the basic LDA, which do not automatically infer the number of topics. This introduces the problem of redundant topics. Adapting a nonparametric bayesian models as Hierarchal Dirichlet Process (HDP) proposed by [Teh et al., 2006] is a possible direction toward the solution of this problem.

Another limitation lays in reading the results of this model. Users need to spend some effort comprehending the ways they can read the model results and maximize the benefit from the model results. A well-designed interactive visualization is curial here to give the user the freedom of
exploring the results in the way they please and help them answer their questions.

Our future work can take different directions, First, enabling the model to infer automatically the number topics is a possible future direction for this work. Another direction is designing and developing an interactive visualization. This visualization will transform this powerful model to a strong visual analytics tool, and it is a critical future work direction to maximize the benefited and the ease of reading of the model results. Another direction for this work is to enable the DSTM to predict topics and their location from unseen documents. This is a direction we explored in the last part of this dissertation.

8.4 Predictive Analysis

The fourth and last part of this dissertation focused on enabling our DSTM for predictive analysis. A great advantage of this approach is overcoming the limitations in [Wang et al., 2012] approach by enabling their model to predict topics locations and overcome the computational intensive approach they used to update the transition parameter, a parameter calculated from the training data and used for prediction.

A natural direction for future work is examining the prediction approach applicability on different domain specific datasets. Some of the applications we will explore are: predicting future research direction from publications archives and predicting an epidemic outbreak from social media datasets e.g. tweeter, Facebook, and blogs. Another direction is integrating sentiment into this framework. This integration will give valuable insights from the predicted events. The result of this modification will be topics, their locations, and the sentiment that surrounded the predicted topics.
Bibliography


