CAN WE FORECAST CONFLICT?

A FRAMEWORK FOR FORECASTING GLOBAL HUMAN SOCIETAL BEHAVIOR USING LATENT NARRATIVE INDICATORS

BY

KALEV H. LEETARU

B.S. Computer Science, University of Illinois at Urbana-Champaign, 2004

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in the Graduate School of Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

2013

ABSTRACT

The ability to successfully forecast impending societal unrest, from riots and protests to assassinations and coups, would fundamentally transform the ability of nations to proactively address instability around the world, intervening before unrest accelerates to conflict or prepositioning assets to enhance preventive activity. It would also enhance the ability of social scientists to quantitatively study the underpinnings of how and why grievances transition from agitated individuals to population-scale physical unrest. Recognizing this potential, the US government has funded research on "conflict early warning" and conflict forecasting for more than 40 years and current unclassified approaches incorporate nearly every imaginable type of data from telephone call records to traffic signals, tribal and cultural linkages to satellite imagery. Yet, current approaches have yielded poor outcomes: one recent study showed that the top models of civil war onset miss 90% of the cases they supposedly explain. At the same time, emerging work in the economics disciplines is finding that new approaches, especially those based on latent linguistic indicators, can offer significant predictive power of future physical behavior.

The information environment around us records not just factual information, but also a rich array of cultural and contextual influences that offer a window into national consciousness. A growing body of literature has shown that measuring the linguistic dimensions of this real–time consciousness can accurately forecast many broad social behaviors, ranging from box office sales to the stock market itself. In fact, the United States intelligence community believes so strongly in the ability of surface-level indicators to forecast future physical unrest more successfully than current approaches, it now has an entire program devoted to such "Open Source Indicators." Yet, few studies have explored the application of these methods to the forecasting of non-economic human societal behavior and have

primarily focused on large-bore events such as militarized disputes, epidemics, and regime change. One of the reasons for this is the lack of high-resolution cross-national longitudinal data on societal conflict equivalent to the daily indicators available in economics research.

This dissertation therefore presents a novel framework for evaluating these new classes of latent-based forecasting measures on high-resolution geographically-enriched quantitative databases of human behavior. To demonstrate this framework, an archive of 4.7 million news articles totaling 1.3 billion words, consisting of the entirety of international news coverage from Agence France Presse, the Associated Press, and Xinhua over the last 30 years, is used to construct a database of more than 29 million global events in over 300 categories using the TABARI coding system and CAMEO event taxonomy, resulting the largest event database created in the academic literature. The framework is then applied to examine the hypothesis of latent forecasting as a classification problem, demonstrating the ability of a simple example-based classifier to not only return potentially actionable forecasts from latent discourse indicators, but to quantitatively model the topical traces of the metanarratives that underlie them. The results of this dissertation demonstrate that this new framework provides a powerful new evaluative environment for exploring the emerging class of latent indicators and modeling approaches and that even rudimentary classification-based models may have significant forecasting potential.

To my parents Hannes and Marilyn and my dogs Calcite and Aita.

ACKNOWLEDGEMENTS

I would like to offer special thanks to my advisor Les Gasser for going so far above the call of duty on so many occasions in helping to bring this all together and for being such an incredible friend and mentor throughout it all, and to my committee, Orville Vernon Burton, Marshall Scott Poole, and John Unsworth for helping make this a reality. Without each of you, this dissertation would never have been possible.

I would like to thank my department and the Irwin, Boyd Rayward, Josie B. Houchens, and University Fellowships for their financial support in allowing me to write this dissertation. I would also like to thank Alan Craig, who first got me interested in the doctorate program at GSLIS. In addition, I would like to thank Reed Elsevier's LexisNexis Group for allowing me to use their news archives for this research.

I would also like to thank Tony Olcott for so many fascinating conversations and all of his insight on the global media and for first recommending that I explore the latent-physical link of the media.

Finally, I would like to thank my parents Hannes and Marilyn and dogs Calcite and Aita for all of their support, encouragement, and long nights.

TABLE OF CONTENTS

CHAPTER 1: A TEMPLATE FOR FORECASTING SOCIETAL BEHAVIOR 1
1.1 A TEMPLATE AND WORKFLOW FOR CREATING SOCIETY-SCALE BEHAVIORAL DATA 1
1.1.1 ACQUISITION OF NEWS MATERIAL 2
1.1.2 CONSTRUCTION OF THE EVENT DATABASE 2
1.1.3 APPLYING THE DATABASE TO CONSTRUCT A LATENT FORECASTING MODEL 4
1.2 LATENT FORECASTING AS A CLASSIFICATION PROBLEM 4
CHAPTER 2: THE SCIENCE OF CONFLICT FORECASTING 6
2.1 A HISTORY OF APPROACHES TO CONFLICT FORECASTING 7
2.1.1 THE SOLITARY EXPERT: HUMAN ASSESSMENT 8
2.1.1.1 Origins/Basis 8
2.1.1.2 Application 8
2.1.1.3 Limitations 9
2.1.2 POOLING EXPERTS FOR CONSENSUS VIEWS: PREDICTION MARKETS 10
2.1.2.1 Origins/Basis 10
2.1.2.2 Application 10
2.1.2.3 Limitations 12
2.1.3 USING DATA TO EXTRAPOLATE PHYSICAL ACTION FROM PAST PHYSICAL ACTION 13
2.1.3.1 Origins/Basis 13
2.1.3.2 Application 14
2.1.3.3 Limitations 15
CHAPTER 3: MOVING FORWARD: ASSESSING AND FORECASTING POPULATION-SCALE BEHAVIOR THROUGH LATENT MEDIA INDICATORS 17
3.1 ATTENTION ECONOMIES: PRODUCTION VERSUS CONSUMPTION 18
3.2 FORECASTING FUTURE PHYSICAL BEHAVIOR FROM PRESENT LATENT EXPRESSION 20
3.2.1 THE IMPORTANCE OF PERCEPTION 20
3.2.2 DISCOURSE AS A FORECASTING METRIC 21
3.2.3 MEASURING EMOTIONS AND BELIEFS 22
3.2.4 EMOTION AND ASSESSING POPULATION-SCALE BEHAVIORS 24
3.3 “BIG DATA” 25
3.3.1 THE TRANSITION TO REALTIME 26

3.3.2 AUTOMATED SOLUTIONS TO DROWNING IN DATA 27
3.4 MEDIA AS CULTURAL PROXY: ASSESSING REMOTE POPULATIONS 27
3.4.1 REGIONAL VARIATION 28
3.4.2 CULTURAL VARIATION 29
3.4.3 EMOTION AND INFORMATION PROCESSING 30
3.4.4 EMOTION WITHOUT NETWORK-CONTEXTUAL KNOWLEDGE 30
3.5 CORE RESEARCH QUESTIONS 31
CHAPTER 4: METHODOLOGY: QUANTIFYING RHETORIC AND REALITY 34
4.1 QUANTIFYING SOCIETY 34
4.2 NEWS SOURCES 38
4.2.1 AGENCE FRANCE PRESSE 40
4.2.2 ASSOCIATED PRESS 43
4.2.3 XINHUA 46
4.2.4 COMPARING THE SOURCES 49
4.2.5 ADDITIONAL POST FILTERING 51
4.3 CODIFYING SOCIETAL BEHAVIOR 52
4.3.1 EXISTING EVENT DATABASES 52
4.3.2 TABARI AND CAMEO 54
4.3.3 PROCESSING PIPELINE 58
4.3.4 EVENT DATABASE 58
CHAPTER 5: FORECASTING BY CLASSIFICATION 62
5.1 MODEL CONSTRUCTION 65
5.1.1 TEXTUAL SURROGATES 68
5.1.2 FEATURE SELECTION AND WEIGHTING 76
5.1.3 ASSESSING ACCURACY 79
5.2 FIRST EXPERIMENT: USING XINHUA TO FORECAST EGYPT 82
5.2.1 FORECASTING HIGH-EVENT DAYS 87
5.2.2 EXPANDING THE EVENT CATEGORIES 89
5.2.3 USING MULTIPLE TEXT DAYS 92
5.2.4 NARROWING TO PROTESTS 93
5.2.5 PEERING INSIDE THE MODELS 95
5.2.6 LEARNING FROM MORE RECENT KNOWLEDGE 101

5.3 EXPANDING TO OTHER SOURCES AND COUNTRIES 110
5.3.1 EXPLORING THE COUNTRIES 112
5.3.2 PEERING INTO THE MODELS: UNDERSTANDING COUNTRY-LEVEL DRIVING FACTORS 121
5.3.3 TESTING CROSS-COUNTRY MODELS 131
5.3.4 REVISITING PROTESTS 133
5.3.5 TESTING WEEKLY FORECASTS 135
CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS 138
6.1 A TEMPLATE FOR TESTING THE LATENT-PHYSICAL LINK 138
6.2 MAJOR FINDINGS 140
6.3 FORECASTING AS SCIENCE AND COMPARING ACCURACY 144
6.4 METHODOLOGICAL LIMITATIONS AND THEIR MITIGATIONS 146
6.4.1 DATASET BOUNDARIES 147
6.4.1.1 Mitigation 147
6.4.2 DATASET ERRORS 147
6.4.2.1 Mitigation 148
6.4.3 ADVERSE MEDIA EFFECTS 148
6.4.3.1 Mitigation 150
6.4.4 CENSORSHIP AND MEDIA INTERFERENCE 150
6.4.4.1 Mitigation 150
6.4.5 CENSORSHIP: BLOCKING MESSAGES 151
6.4.5.1 Mitigation 152
6.4.6 CENSORSHIP: SATURATING MESSAGES: ORGANIC VERSUS INORGANIC CAMPAIGNS 153
6.4.6.1 Mitigation 153
6.4.7 CENSORSHIP: SELF-CENSORSHIP: HARASSMENT OF THE MEDIA 154
6.4.7.1 Mitigation 154
6.4.8 THE PARADOX OF PREDICTION AND SELF-FULFILLING PROPHECIES 155
6.4.8.1 Mitigation 156
6.5 CONTRIBUTION TO THE LITERATURE 156
6.5.1 NATIONAL SECURITY AND CONFLICT PREVENTION 157
6.5.2 A NEW EMPIRICAL TESTBED FOR THE SOCIAL SCIENCES AND HUMANITIES 157
REFERENCES 160

CHAPTER 1: A TEMPLATE FOR FORECASTING SOCIETAL BEHAVIOR

Over the past decade there has been a growing literature incorporating latent linguistic-based indicators into statistical models that attempt to forecast the risk of future societal unrest. A common theme to emerge from this literature is the lack of widely-available cross-national high-resolution catalogs of human behavior that would allow such work to move beyond studying high-mortality episodes like interstate war, epidemics, and natural disasters towards their small-bore precursors like individual protests and peace appeals. Training and testing such forecasting models requires quantitative databases of behavior that codify a humanistic concept like a "protest" into its measurable attributes of location, date, and groups involved. This allows the construction of time series sequences that codify the conflict and cooperation behavior of a nation over time which can be aligned with latent indicators of interest to explore possible predictive relationships, such as time-lagged correlations.

1.1 A TEMPLATE AND WORKFLOW FOR CREATING SOCIETY-SCALE BEHAVIORAL DATA

This dissertation presents one possible approach to constructing such small-bore society-scale behavioral data on an ad-hoc basis through a computational template and workflow for evaluating political forecasting models based on latent linguistic indicators. It covers the data acquisition process, the application of specialized software to convert that data into a quantitative database of discrete activities, and the utilization of that database to construct a latent forecasting model. One of the most powerful aspects of this template and workflow is that it does not require extensive human or computational resources and can be performed by a single individual using a standard desktop computer. This enables its application to ad-hoc or resource-constrained analysis, such as use within

classroom settings for teaching or doctoral research, whereas previous efforts required large human teams and institutional computing resources, placing them out of reach of such applications.

1.1.1 ACQUISITION OF NEWS MATERIAL

The mainstream news media has been one of the primary sources for the cross-national study of human political behavior due to its widespread availability, longitudinal and cross-national coverage, and emphasis on the conflict and cooperation actions that underlie many types of societal-scale behavior. The first stage of the template presented in this dissertation is therefore a workflow for downloading mainstream news coverage in machine-readable electronic format from the LexisNexis Academic Universe service. This news aggregator archives more than 5,000 international, national, regional, and local mainstream news sources covering nearly forty years. Case studies are presented of the three largest international news agencies: Agence France Presse, Associated Press, and Xinhua. For each of these three newswires, this dissertation's workflow is used to evaluate the source, determine the necessary Boolean query to filter irrelevant articles, and assess its geographic and temporal coverage. At the end of this stage of the workflow, the necessary news material has been assembled, filtered to exclude the majority of irrelevant content, and prepared for processing.

1.1.2 CONSTRUCTION OF THE EVENT DATABASE

Once the necessary mainstream news articles have been downloaded from LexisNexis, the next section of the workflow demonstrates the conversion of that content into a quantitative database of small-bore political events using the open source TABARI software package. The TABARI program is a fully automated event coding system that operates in unattended batch mode, accepting as input a set of files containing news articles and outputting a set of tab-delimited files listing all of the "events" it identified within that coverage. Due to computing limitations, TABARI was historically applied only to

the first sentence of each news article, but recent literature has recommended operating it in its extended mode to process the full text of each article to capture secondary events, which is done here. In addition, TABARI's ability to construct georeferenced events, in which events include information about the specific city the activity took place in, rather than only recording the country, is demonstrated. This ensures that the system properly records the intensity of unrest in high-conflict areas, such as the ongoing Syrian civil war, and provides a spatial dimension to the event database that could be used in future research.

TABARI places each event into the CAMEO event taxonomy, which defines more than 300 types of events in 20 major categories. These events can also be reduced to a more compact set of "Quad Classes" in which events fall under the headings of Material Conflict, Material Cooperation, Verbal Conflict, or Verbal Cooperation. Since many of the more specific CAMEO event types, such as those dealing with the use of weapons of mass destruction, usually capture relatively few instances, the coarser Quad Class headings allow easier access to the data. Examples are provided analyzing the dataset at both the full-resolution and Quad Class levels.

TABARI and its CAMEO taxonomy are the most widely-used tools for constructing event databases in the quantitative political science literature. They are also used in the United States Department of Defense ICEWS system, which catalogs global political behavior each day for the US military and TABARI was the only system to pass the rigorous accuracy benchmarks assessed during the construction of ICEWS. TABARI's fully automated operation means it can be used to process millions of articles, including the entirety of multiple newswires, enabling the comparison of how news outlets differ in their coverage of major events, which is explored in Chapter 5.

1.1.3 APPLYING THE DATABASE TO CONSTRUCT A LATENT FORECASTING MODEL

Chapter 4 demonstrates using this workflow to download all international coverage from the Agence France Presse, Associated Press, and Xinhua news wires through LexisNexis Academic Universe and applying the TABARI package to identify 29 million events in the CAMEO taxonomy from this content. Chapter 5 then utilizes this database to construct a series of basic latent forecasting models that identify word usage patterns that forecast the likelihood of various classes of events occurring in the future. Here, the event database is used to construct a time series sequence that counts the total number of events recorded in the database each day to create a binary variable classifying each day as "High Event" or "Low Event."

Historically, latent forecasting models have used one set of data for their input latent measures and a different dataset for the list of events to be forecasted. This makes assessment of their accuracy difficult in that it is difficult to determine whether a poorly-performing model is due to limitations of that particular modeling approach or whether it is due to a mismatch between the types of events discussed in the news material used for latent input and the source material used to construct the event database. TABARI allows an event database to be constructed from the same news content used for the latent input, eliminating this possible source of error.

1.2 LATENT FORECASTING AS A CLASSIFICATION PROBLEM

Chapter 5 leverages this template and its workflow to explore the hypothesis of forecasting as a classification problem, providing an example of using the template to evaluate a latent forecasting model. Here, all news coverage published in a given day mentioning a country is concatenated to construct a "document" and the categorical label assigned to that "document" is "High Event" or "Low

Event," determined by the number of events recorded in the event database as occurring the following day. This is then examined using a standard Naïve Bayesian text classification software package to evaluate the accuracy that different subsets of the article text provide in forecasting future behavior. The resulting Naïve Bayesian models that offer measurable forecasting accuracy are then dissembled and their conditional probability term tables examined to demonstrate the use of such models to capture topical traces of the underlying metanarratives of each country. Here, forecasting is used as a tool to codify linguistic indicators that appear more frequently before certain types of events and less frequently when those events do not subsequently occur, rather than focusing purely on creating operational forecasts. Five countries are examined, representing each region of the world, and are explored through the respective coverage of three major newswires.

In conclusion, this dissertation demonstrates an end-to-end template and associated workflow for acquiring large volumes of electronic mainstream news media, computationally assembling a quantitative event database from that content, and applying that database to test a proposed latent forecasting modeling approach.

CHAPTER 2: THE SCIENCE OF CONFLICT FORECASTING

The ability to successfully forecast impending societal unrest, from riots and protests to assassinations and coups, would significantly increase the ability of nations to proactively address instability around the world, intervening before unrest accelerates to conflict or prepositioning assets to enhance preventive activity. It would also enhance the ability of social scientists to quantitatively study the underpinnings of how and why grievances transition from individual latent agitation to population-scale physical unrest. Recognizing this potential, the US government has funded research on "conflict early warning" and conflict forecasting for more than 40 years and current unclassified approaches incorporate a significant diversity of data sources from telephone call records to traffic signals, tribal and cultural linkages to satellite imagery ("Science of Civil War", 2012).

Much of this new research is being driven by a new-found acceptance within the academic community of scholarly research on conflict that may be government-funded or have operational utility. The US Army's first major foray into the academic partnership for forecasting conflict debuted in 1964 with Project Camelot. Camelot was an attempt to "use the systems analysis approach to studying … internal conflict and insurgency", addressing the "very incomplete knowledge and understanding in depth of the internal cultural, economic, and political conditions that generate conflict between national groups" (Lowe, 1966). Its primarily goal was to bring together "sociologists, anthropologists, economists, psychologists, geographers, and other social science specialists" to conduct a "feasibility study to see how far it is possible at the present time to push the state of the behavior science and technology toward increasing our capability to anticipate social breakdown and to suggest remedies" (Lowe, 1966). Yet, the ensuing backlash led to the project being canceled less than a year later, among commentary that such research was "a grave violation of the most elementary rules of professional ethics which must

govern relations between scientists" and "may endanger the development of sociology." In contrast, 47 years later, a vast array of data-driven social science efforts endeavor to "reveal sociological laws of human behavior – enabling [researchers] to predict political crises, revolutions and other forms of social and economic instability, just as physicists and chemists can predict natural phenomena" (Markoff, 2011).

Yet, current approaches have yielded relatively poor outcomes: one recent study showed that the top models of civil war onset miss 90% of the cases they supposedly explain (Ward et al, 2010). At the same time, emerging work in the economics disciplines suggests that new approaches, such as the use of media-derived latent indicators, may offer greater predictive power of future physical behavior. Further complicating matters, there are differing conceptions of what it means to "forecast" conflict. This can range from predicting the specific day and location of an individual riot (IARPA, 2011) to providing "risk" estimates that assess the broad likelihood of a particular class of event (O'Brien, 2010). Yet no matter the mechanism or conceptual approach, the ability to successfully forecast future societal unrest with a long enough time horizon to take action would greatly enhance the nation's ability to mitigate or moderate conflict and offer social scientists and humanists a wealth of new experimental data on the underpinnings of societal strife.

2.1 A HISTORY OF APPROACHES TO CONFLICT FORECASTING

Societies have long attempted to forecast where future unrest may occur in order to maintain stability. Scholars have similarly sought to identify the processes through which unrest solidifies into conflict to better their understanding of how societies function. Historically, approaches to conflict forecasting have taken three distinct forms: individual experts synthesizing available information and rendering a

judgment, collections of such experts pooled together into a consensus view, and data-driven computer models that use patterns of past physical behavior to predict future physical behavior. This section introduces these approaches to conflict forecasting, describing their applications and limitations. It also demonstrates some of the significant difficulties encountered when attempting to make judgments of future behavior.

2.1.1 THE SOLITARY EXPERT: HUMAN ASSESSMENT

2.1.1.1 Origins/Basis

The oldest and most widely-utilized mechanism for assessing stability and forecasting future unrest is the use of human "experts" who synthesize available environmental information through the lens of their own experiences and beliefs to develop intuitive judgments on how likely a given outcome of interest may be. The benefits of using experts is obvious: they are relatively cheap compared with alternatives, and intuitively, an individual with significant experience studying an area of interest over many years might seem more likely than a mathematical equation to be able to predict where unrest is likely to break out next (Lowe, 1966).

2.1.1.2 Application

The majority of formal intelligence assessment of state fragility and likelihood of conflict is based primarily on subject matter experts and intelligence analysts synthesizing available information (Olcott, 2012). In addition, most annual "state fragility" indexes incorporate expert judgments in some form, either directly as one of their input variables, or indirectly, by incorporating variables from datasets that

are based on expert assessments. Examples include the Fund for Peace's annual Failed States Index (n.d.), Polity's State Fragility Matrix (n.d.), and USAID's State Fragility Index (n.d.).

2.1.1.3 Limitations

It takes only a quick glance at many of the public fragility indexes from 2010 to note that many of the countries that have collapsed in the last several years were given only a low probability of serious unrest. At the same time, many of the countries listed at greatest risk of collapse in 2010 have experienced no change in leadership or substantial increase in civil unrest. Therein lays one of the major limitations of expert assessments: they tend to be wrong more often than they are right. Indeed, if expert assessments were able to forecast unrest at operationally-acceptable accuracy levels, there would likely be little interest in the vast array of new forecasting approaches being funded by the US government today.

The reasons for flawed forecasts are myriad, but given the significant foreign policy stakes inherent in conflict forecasting, and the personalities of the policy leaders for whom such forecasts are designed, political considerations can have an outsized influence on driving outcomes. Sherman Kent, one of the founders of Western intelligence, noted in 1949 that the ultimate consumer of such forecasts, policymakers, considered themselves to be "experts" on all issues and would absorb new information (such as forecasts) only "to the degree that its findings coincided with that [policymaker's] previous understandings" (Kent, 1949 as cited in Olcott, 2009). Indeed, Kent believed that the intelligence community must be kept separate from policymaking to prevent analysts from "swinging behind the 'policy' of the operating unit," actively altering their judgments to support the desired policy decisions of their employing policymakers (Kent, 1949 as cited in Olcott, 2009). Thus, early efforts at integrating

data-driven “modern social science research techniques” were viewed as “conflict[ing] with [this] traditional ‘intuitive’ approach to foreign policy based on ‘total immersion’ into the affairs of a country or area” (Lowe, 1966).

2.1.2 POOLING EXPERTS FOR CONSENSUS VIEWS: PREDICTION MARKETS

2.1.2.1 Origins/Basis

Another limitation of expert-based forecasts is their variance. A collection of experts may each offer significantly different estimates based on their varied backgrounds, assumptions, and knowledge (Weigle, 2007). One common mechanism for working around the limitations of individual experts is to pool their judgments together to aggregate the collective wisdom of hundreds or even thousands of experts from an array of backgrounds, disciplines, and motives. For example, IARPA’s Aggregative Contingent Estimation (ACE) program aimed to “dramatically enhance the accuracy, precision, and timeliness of forecasts for a broad range of event types, through the development of advanced techniques that elicit, weight, and combine the judgments of many intelligence analysts” (IARPA, 2010).

2.1.2.2 Application

The Political Instability Task Force (PITF) was an attempt begun in 1994 by the US Government to revive the ideals of Project Camelot and bring together a collection of the leading scholars across the nation to “assess and explain the vulnerability of states around the world to political instability and state failure” and “develop statistical models that can accurately assess countries' prospects for major political change and can identify key risk factors of interest to US policymakers” (“Political Instability”, online). Rather

than relying on the predictions of a single expert, PITF was developed to bring together into a single working group the views of noted researchers across the country from a range of academic disciplines to generate a more comprehensive and cohesive view of conflict.

A more formalized implicit instantiation of such expert aggregation is the so-called "prediction market," which is "essentially [a] speculative marke[t] created for the purpose of aggregating information and making predictions" (Williams, 2012). In a traditional expert aggregation model, experts formalize their forecasts into explicit statements, which are discussed and compared. In the prediction market, the outcome to be forecasted is treated as a tradable commodity in which pricing is based on traders' beliefs in the probabilities of various versions of that outcome. Under this model, experts use their beliefs to trade in this marketplace in an implicit process.

In May 2001 the Defense Advanced Research Projects Agency (DARPA) announced a new initiative to leverage prediction markets for forecasting civil conflict. The Electronic Market-Based Decision Support (DARPA SB012-012) was designed to aggregate expert assessments of the "analysis of the likelihood of events that motivate the Quadrennial Defense Review, prediction of the timing and impact on national security of emerging technologies, analysis of the outcomes of advanced technology programs, or other future events of interest to the DOD" (DARPA, 2001). This evolved into the FutureMAP (Futures Markets Applied to Prediction) program designed to "identify the types of market-based mechanisms that are most suitable to aggregate information in the defense context … develop information systems to manage the markets, and … measure the effectiveness of markets for several tasks" (DARPA, 2002). Like its predecessor, this program aimed to pool expert assessments to forecast "political stability in regions of the world, prediction of the timing and impact on national security of emerging technologies,

analysis of the outcomes of advanced technology programs, or other future events of interest to the DoD" (DARPA, 2002).

A key focus of the market system was the combination of broad assessments from experts familiar with a given outcome of interest, but also the ability of such prediction markets to provide early warning of unexpected outcomes based on "the rapid reaction of markets to knowledge held by only a few participants" (DARPA, 2002). This reflects the importance of a few knowledgeable actors over the broad assessments of traditional experts who may or may not have specific knowledge regarding the outcome of interest. The program was intended to operate two exchanges, one restricted to 2,000 US Government employees designed to test pooling of Government intelligence and one with 6,000 traders that would include "academia, the media, policy research institutes, insurance companies, non-governmental organizations and the general public" (Weigle, 2007).

2.1.2.3 Limitations

Due to public outrage over the notion of the US Government facilitating private individuals profiting on tragedies such as assassinations and wars, the program was ultimately canceled before it began, so no results were obtained on the accuracy of such approaches for conflict forecasting. Public perception of such approaches makes it unlikely that such programs will be permitted in the foreseeable future. In addition, an increasing number of high-profile inaccurate predictions have resulted in expanded discussion of the limitations of the prediction market approach (Leonhardt, 2012). Since markets require traders to take action to participate (such as placing purchase orders), the pool of users is small, decreasing the likelihood that any one of them is a knowledgeable insider on the issue to be forecasted. This was most apparent in the 2012 United States Supreme Court battle over the Obama

administration's Affordable Health Care for America (HR 3962), when one of the major prediction markets offered a 75% probability that the law would be ruled unconstitutional (Leonhardt, 2012). Yet, despite the inaccuracy of the prediction market, there did appear to be publically-accessible early warning signs: "rumors began to circulate in Washington… liberals around town who might have reason to know the outcome seemed happy… and a couple of conservative justices had seemed angry when the court met three days before the announcement" (Leonhardt, 2012).

Thus, while prediction markets attempt to work around the limitations of individual experts by aggregating a much larger collective knowledgebase, they still require explicit action to participate and thus represent a small pool of users. On the other hand, even in the case of a Supreme Court ruling where only a few elites knew the outcome beforehand, early warning indicators manifested themselves in open forums such as media coverage. This suggests that passively pooling assessments via the media over a vastly larger population, rather than actively polling a small number of self-selected users, could potentially yield better accuracy (Leonhardt, 2012). This is a concept that will be returned to in the following chapter.

2.1.3 USING DATA TO EXTRAPOLATE PHYSICAL ACTION FROM PAST PHYSICAL ACTION

2.1.3.1 Origins/Basis

While expert assessments and prediction markets both incorporate available situational knowledge, they do so only through the interpretive lens of participating experts. Information that conflicts with the expert's assumptions or hypotheses may be discounted and only a limited subset of available information may be incorporated into the final assessment (Kent, 1949 as cited in Olcott, 2009). Data-

driven models therefore seek to utilize quantitative statistical methods and large-scale data analysis to derive computational equations that are dependent only on the underlying patterns in the data. Human interpretation still plays a role in that humans select the statistical techniques and input data used to derive the models, but the forecasts themselves are based solely on extracted quantitative patterns from the input data, rather than the amorphous intuition of an expert or pool of experts ("Science of Civil War", 2012). Forecasts can thus be deconstructed to explore their entire reasoning path and provenance, something that is difficult with human judgment. The underlying concept of these models is therefore to observe the past and present states of the physical world, identify patterns in physical behavior, and use those patterns to forecast future physical activity.

2.1.3.2 Application

Physically-based models can yield successful results when applied to macro-social-level phenomena like population-scale migration in the aftermath of a natural disaster. A recent study of the 2010 Haitian earthquake (Lu et al, 2012) using individual-level cellular records found that when a natural disaster occurred, residents in the affected area dispersed back to the locations they were in during the previous Christmas holiday. By snapshotting the location of all cellphones at Christmas, immediately before the earthquake, and immediately afterwards, the authors found that the places people traveled to during the holidays were the same places they returned to in a disaster. This captured that Christmas is an inherently family-oriented time of the year when families come together, with children returning to their parent's home, or distant relatives all congregating at one another's houses.

Such national-scale migration patterns could also be identified through government records and previous address lists, linking children to their parents and relatives to one another. However,

government records ordinarily capture only familial ties, not friendships. During the Christmas season, people tend to visit the homes of close friends and coworkers, with cellular data capturing these non-familial links that are traditionally inaccessible to analysis. In the case of this large-scale natural disaster, where family homes were often within the disaster area, the cellular data allowed researchers to map these secondary refuges. In this way, high-resolution observational locative data like cellular records may permit the mapping of friendship ties in ways that can offer household-level forecasts of migration after a major disaster.

In the realm of conflict forecasting, one of the primary measures of conflict used to populate modeling approaches is the so-called "event database." Such databases are essentially daily or annual catalogs of riots, protests, assassinations, violent attacks, and other incidents of physical unrest in a given country, used to look for temporal patterns in the timing and intensity of the records. Databases like the European Media Monitor (EMM) (Atkinson & Van der Goot, 2009) and DARPA's Integrated Conflict Early Warning System (ICEWS), (O'Brien, 2010) described in more detail in Chapter 4, compile large archives of incidents of physical unrest in countries of interest in realtime, producing regular reports summarizing major emerging spatial-temporal patterns. The contents of these event databases may be compared against a portfolio of institutional characteristics such as GDP and infant mortality ("Political Instability", online), or ethnic fragmentation and democracy level (Hegre & Sambanis, 2006), or explored solely in the temporal dimension, leveraging autocorrelation signatures (Shrodt, 2000).

2.1.3.3 Limitations

When transitioning from predicting disaster-related population movement to predicting human conflict, current approaches have yielded lower accuracy, even when focusing on whole-country violence. For

example, a recent survey of the literature on the most-cited models of civil conflict found that those models fail to forecast 90-100% of whole-country conflict, with one model correctly forecasting 0 of 107 wars it claimed to explain (Ward et al, 2010). Even the most advanced models using sophisticated statistical techniques such as Hidden Markov Models exhibit false positive rates exceeding 70% and false negative rates of greater than 10% (Shrodt, 2000). This is potentially due to the fact that human behavior is highly irregular, which complicates using only past behavior to forecast future behavior. A recent study applying random fractal theory analysis found that a wide array of social phenomena exhibit "non-stationary, on-off intermittent or Levy walk processes" making internal forecasting extremely difficult (Gao et al, 2012). This suggests that purely behavioral forecasting may need to be supplemented by the integration of secondary latent dimensions that attempt to proxy societal-scale beliefs about the environment, which will be explored in the next chapter.

Despite the US government spending more than $125 million over the last four years on physically-based forecasting models, including $38 million on ICEWS, current physically-based forecasting efforts have exhibited a relatively poor track record. As a recent Wired article noted, "we do better than human efforts, but not by much" and "all of our models are bad, some are less bad than others" (Shachtman, 2011). Forecasting experiments in the ICEWS project correctly predicted just a quarter of catastrophic whole-country violence (O'Brien, 2010). Most notably, current efforts tend to surface only simmering conflicts with long slow progressions towards unrest (which human experts tend to do well at picking up and which leave ample room for military and diplomatic planning), but not the rapid highly fluid "sudden" conflicts that can have the greatest policy implications (Shrodt, 2000).

CHAPTER 3: MOVING FORWARD: ASSESSING AND FORECASTING POPULATION-SCALE BEHAVIOR THROUGH LATENT MEDIA INDICATORS

Of the numerous approaches to forecasting human conflict, perhaps one of the most promising is the emerging measurement of latent indicators from written media expression to assess the potential trajectory of future physical behavior. Each day "people from Bangladesh to Buenos Aires busily tell one another and their neighbors what they see, what they think, and what is important to them … offering unparalleled visibility into what global society is paying attention to" (Leetaru & Olcott, 2012). Through enabling the sharing of their "most intimate thoughts, actions, and likes or dislikes for the world to see" the media is "allow[ing] faster, more fine-grained, round-the-clock access to societal reaction around the globe" (Leetaru & Olcott, 2012). The "constant stream of daily life that flows across [media] platforms provides rich contextual background information on the narratives of each region and culture" (Leetaru & Olcott, 2012). In essence "Citizens are becoming a vast ground-based social sensor network, providing a continuous real-time picture of almost every corner of the world" (Leetaru & Olcott, 2012) and the speed and scope of this sensor network are increasing by the day (Toffler, 1984).

The mainstream news media has long been a data source for the study of conflict, using both discrete event records and latent indicators. The level of "attention" received by a country has been used as a proxy of its "importance" (Woolley, 2000), while the "stability" or "instability" of a given country is measured through the creation of databases of discrete "event" entries. Conflict incidents such as riots, protests, and attacks, and cooperative actions such as promises of aid or peace appeals are aggregated

and scored to generate a timeline of the intensity of conflict and cooperation within a nation (Gerner & Schrodt, 1996).

3.1 ATTENTION ECONOMIES: PRODUCTION VERSUS CONSUMPTION

In actuality, the true question of concern in assessing population behavior is not what populations are saying, but rather what populations are concerned about and paying attention to (Olcott, 2010). Human beings are highly limited in the volume of information they can physically process, creating significant competition for limited mental resources in an information-rich environment. The study of "attention economies" explores the ways in which humans balance these conflicting information demands and the decision-making process through which any given item of information is brought forward to conscious or subconscious awareness (Simon, 1971; Davenport & Beck, 2001; Toffler, 1984). From the standpoint of the news media, a given person can consume just a fraction of the vast array of novel stories published each day. It is therefore far more informative to monitor just the stories a population is reading and attending to, rather than examining every article published in a given day.

As Leetaru & Olcott (2012) note, media analysis was originally "developed during World War II and the Cold War as a surrogate for leadership analysis, created to use state-controlled newspapers and other state media as the only available means to study the perceptions and intentions of leaders and elites in areas about which we had no other sources of information." They argue further that "the reason that method worked was, in retrospect, an accident of technology—it was far cheaper to receive information (buy a newspaper, purchase a radio receiver) than it was to create and send it (publish a newspaper, own a radio studio or TV station)" and therefore "even in states in which media were not state

controlled, they still represented the interests, and viewpoints, of the elites" and thus at the very least captured what "the elites wanted the masses to see, hear, and think" (Leetaru & Olcott, 2012).

In contrast "the most important indicator for understanding other people is to learn what they are interested in and what they pay their attention to—this permits us to understand their hopes, their fears, their aspirations, and their value systems" (Leetaru & Olcott, 2012). Electronic media offer new mechanisms for measuring this "attention economy" (Toffler, 1984; Olcott, 2012). In the past researchers could measure message uptake only at the level of the medium itself, such as how many copies of a newspaper were purchased, but could make no assumptions about whether any given article within that paper was read. In the era of web-based news distribution, each individual story has its own distinct identifier and is accessed individually, allowing the measurement of actual readership to the level of the story. This resolution of analysis is increasingly being used to rank reporters by the popularity of their articles (Peters, 2010) and by academic researchers to better understand the stories that most resonate with the public (Tierney, 2010).

Unfortunately, while the digital era offers the unique ability to ascertain message-level consumption, such data is closely held by publishers. Message production is therefore still the only widely available measure available for external analysis and will be the focus of this dissertation. However, it is useful from a methodological standpoint to note that it is attention that likely provides the most accurate insight into societal behavior, and that the production measures that are available for analysis are merely a proxy for this consumptive behavior.

3.2 FORECASTING FUTURE PHYSICAL BEHAVIOR FROM PRESENT LATENT EXPRESSION

The availability of new technology and datasets, discussed in more detail later in this chapter, are enabling the deeper exploration of the latent dimensions of narrative discourse, treating the information environment not as a collection of facts, but as a collection of views onto those facts. This is the basis of formalized Open Source Intelligence (OSINT), founded more than 70 years ago to use the tone and thematic content of global news media to offer insights into world events, including the risk of conflict (Roop, 1969; Mercado, 2001). Physical manifestations of unrest, while often described as "sudden" are in fact the result of a long process of emotional discontent that eventually crosses a threshold to physical action, as in labor strikes (Thompson & Borglum, 1973). In fact, a significant literature studies the concept of these "tipping points" that transition populations from simmering dissatisfaction to physical collective action (Schelling, 1978; Melucci, 1996; Marwell & Oliver, 1993; Oliver, 1993; Yin, 1996). For example, while mass crowds of protesters appeared suddenly during the January 2011 Egyptian revolution, frustration with the government had simmered for a long time, gradually growing until it reached a crossover point when the population took to the streets in a physical manifestation (Stolberg, 2011).

3.2.1 THE IMPORTANCE OF PERCEPTION

Perception of the environment and of self can play a far more powerful role than the actual factual status of that environment. Views of self and group membership can substantially influence how situational information is processed (Ellemers, 2012), while the degree to which an issue matches internal narratives can define the flexibility towards negotiation of goals relating to that issue (Atran & Ginges, 2012). Even subtle shifts in how an issue is framed can sometimes mitigate conflict around that

issue. Such in-group/out-group language leaves measurable residues on the media narrative around those issues that can be measured through automated approaches. For example, Althaus & Leetaru (2011) found that global mainstream media coverage of the 2003 US invasion of Iraq, as captured by the CIA Foreign Broadcast Information Service (FBIS), exhibited strong country-level traces of in-group/out-group linguistic indicators that mirrored their relationships to the conflict.

3.2.2 DISCOURSE AS A FORECASTING METRIC

Emotion has gained traction as a method for operationalizing the forecasting of human behavior because it requires only the identification of surface patterns in observational data, not the creation of theories to explain those patterns. Despite the availability of new data, no theories of conflict have managed to hold across all circumstances. For example, each individual in the crowd of protesters in Egypt's Tahrir Square in January 2011, while united against their oppressive government, likely had a different set of specific triggers that caused him or her to transition from latent unrest to joining the physical protest movement (Fahim et al, 2011). It would simply be impossible at present to attempt to capture that level of complexity (and that level of data is simply not available), and thus the idea behind latent forecasting is that surface-level indicators synthesize this complexity into a single numeric indicator, allowing one to focus on empirical patterns rather than theoretical descriptions of those patterns (Bollen et al, 2011).

Instead of directly observing the 88 different variables that have been used in physically-based conflict models (Hegre & Sambanis, 2006), latent media indicators are designed to proxy how populations are synthesizing, interpreting, and reacting to the effect of all available knowledge. In addition, knowledge dissemination occurring through non-electronic channels, such as face-to-face communication, is

reflected in the latent views expressed through this environment. In essence, each individual is acquiring, filtering, synthesizing, and disseminating information through a myriad of channels and modalities both online and offline, but the end result of all of those inputs on the individual's emotional state and potential future actions may be proxied at least in part through these latent indicators (Bollen et al, 2011; Golder & Macy, 2011; Leonhardt, 2012; Gruhl et al, 2004; Olcott, 2012; Bean, 2011). This is one of the forces driving the exploration of latent measures, as it suggests there is not the need to model populations at the individual level or to begin with theoretical understandings of conflict motivation that may not hold across cultures or time periods. Finally, these indicators are available in a realtime stream, rather than the annual updates of most indicators, meaning they can adjust instantly to exogenous shocks, such as natural disasters.

3.2.3 MEASURING EMOTIONS AND BELIEFS

Emotional responses and beliefs can be measured through a range of mechanisms including in-person field surveys, such as Gallup's poll of worldwide negativity (Clifton, 2012), highly specialized brain scans (Ellemers, 2012), or through automated assessment of open media data (Roop, 1969; Olcott, 2012; Bean, 2011). Field surveys offer the most flexible collection method in that interviewers may ask any question of interest and can adjust the survey questions based on the responses. However, since they require physically visiting or calling each respondent, field surveys are extremely expensive and slow, and as later sections will detail, may yield a postured worldview designed to appear respectful of authority, rather than a genuine worldview. Brain scans potentially offer the truest view of a person's actual thoughts and beliefs (Nishimoto et al, 2011; Kay et al, 2008; Ellemers, 2012), but require extremely specialized equipment and medical facilities and respondents must travel to the brain scanner and undergo a preparation regime in some cases.

Computerized methods have therefore become the dominant mechanism due to their greater scalability. Automated measures of latent indicators assess emotional or thematic dimensions of text by measuring the density of predefined lists of words known as lexicons. There is a wide array of such dictionaries: Leetaru (2011) explored more than 1,500 dimensions from seven different collections including the Regressive Imagery Dictionary (Martindale, 1975), WordStat's version of WordNet 2.0 (Provalis Research, 2005) and the 1911 *Roget Thesaurus* (Provalis Research, 2003), both the H4 and Lasswell General Inquirer dictionaries (Stone, *et al.*, 1966), the Body Type Dictionary (Wilson, 2006), and the Forest Value Dictionary (Bengston and Xu, 1995). Such lexicons are typically applied to mainstream and social media, which capture a snapshot of the real–time public information environment (Stierholz, 2008). News contains far more than just factual details: an array of cultural and contextual influences impact how events are framed for an outlet's audience, offering a window into national consciousness (Gerbner and Marvanyi, 1977). A growing body of work has shown that measuring the "tone" of this real–time consciousness may accurately forecast many broad social behaviors, ranging from box office sales (Mishne and Glance, 2006) to the stock market (Bollen et al, 2011). However, the predictive power of emotional dimensions appears to be heavily context dependent: Bollen et al (2011) found that calm versus anxious language had the greatest predictive power over stock market movements, while Leetaru (2011) found that positive versus negative language offered the greatest insight into future country stability and Chadefaux (2012) found it was conflict-related language itself that had the greatest predictive power.

3.2.4 EMOTION AND ASSESSING POPULATION-SCALE BEHAVIORS

Several studies funded by the United Nations in the last two years have explored the ability of latent media indicators to offer enhanced insight into potential conflict early warning signals. A 2011 study (Crimson Hexagon, 2011) used English and Bahasa tweets to measure population-level concerns around "food, fuel, finance, and housing in the US and Indonesia." It found that thematic data derived from Twitter matched the ground reality (such as tweets on the price of rice matching food price inflation statistics) and that Twitter acts as a useful remote proxy for the thematic concerns of a population. Both broad concerns and specific immediate needs like food are represented in the Twitter stream. However, it also found that Twitter primarily reflected immediate needs and concerns and was a poor indicator of "long term aspirations." Specifically the authors found that Twitter was most useful for "now-casting" and that Twitter "primarily sheds light on the perceptions of the moment, and may be less suited to understanding how people perceive the future," yet this conclusion was largely anecdotal, as they did not derive strong statistical findings to support it.

Another UN-funded study, also from 2011 (SAS, 2011), examined the unemployment-related discourse of half a million blogs, forums, and news websites over two years in the US and Ireland. The authors computed a range of thematic and emotional indicators (the authors allude to these indicators in general terms but do not provide significant technical information on their measures) and ran cross-correlation tests to determine which ones aligned with lagged unemployment data. They found that rises in the intensity of "confused" emotion led the unemployment rate by three months, while the intensity of thematic discussions of housing loss spiked two months after increases in unemployment. Transportation discourse was also seen to peak one month ahead of unemployment, indicating an increased demand for public transportation. The authors show that both thematic and emotional scores

provide forecasting capacity of future unemployment trends. In the United States, both Hostile and Depressed moods increase four months before unemployment rates spike, Uncertainty increases with unemployment, and Housing Loss increases two months after unemployment increases and Auto Repossession three months after. They also found measurable differences between the United States and Ireland in the indicators that offered the greatest predictive power: for example, Hostile and Depressed language peaks in the United States prior to unemployment spikes, while Anxious and Confused language peaks in Ireland and Confident language decreases.

The first study, relying solely on Twitter data, found latent media indicators to be universal, but more of a "now-cast" than a future forecasting tool. The second, incorporating a wider portfolio of media sources including mainstream media, found much stronger forecasting capacity, but also differences across countries. This captures one of the greatest limitations of current research in that there are only a small number of studies available in the open literature and they have largely focused on relatively narrow topics in just a handful of geographic locations. However, one emerging theme from the literature is that by examining multiple media outlets, both local and international, rather than relying on a single source, and using mass correlation of a wide range of indicators against a specific target variable, as opposed to studying only those measures theoretically suggested, latent indicators may offer measurable predictive insight.

3.3 "BIG DATA"

Perhaps the greatest trend driving the ability to remotely assess populations through the media is the so-called "big data revolution" that is providing the datasets, statistical techniques, software algorithms,

and hardware platforms that are enabling improved approaches to studying human society at population scale. The past few decades have witnessed significant changes in the quantity and availability of information on human society. A recent study by IBM (Cha, 2012) pegged the total volume of new data created by humans each day at more than 2.5 quintillion bytes, while more than a quarter-billion photographs are uploaded to Facebook daily by over 845 million active users connected by more than 100 billion friendship links (Facebook, 2012). The volume and velocity of this digital record is enabling computational study of phenomena previously examined primarily by hand, allowing scholars to record change at far more precise time scales (Eisenstein et al, 2010).

3.3.1 THE TRANSITION TO REALTIME

Simultaneously, there is an increasing demand for data to be delivered in realtime. One of the driving forces behind automated media assessment is that while ground-based surveys may offer greater flexibility in the set of questions that may be asked, their cost and logistical constraints mean data are not available for weeks to months, rather than streaming in every few seconds with media analysis. Even Bloomberg's ranking of the world's richest has moved from an annual to a daily ranking to meet increasing demand for realtime updates, applying automated algorithms to its vast economic indicators (Halzack, 2012). The United States Geological Survey recently integrated Twitter monitoring into its earthquake alert service to provide a realtime indicator of "human impact" after it found that 90% of earthquake-related tweets are actual reports of ongoing earthquakes (Meier, 2012).

3.3.2 AUTOMATED SOLUTIONS TO DROWNING IN DATA

The notion of incorporating more information arriving at a faster rate is almost antithetical to the human-driven analytical mindset of intelligence. More than half a century ago in 1949 the intelligence community was already coping with a "flood of information" (Olcott, 2010) and by 1966 the CIA Inspector General's "Cunningham Report" noted that the intelligence community was acquiring "too much information and that, failing to get important information, it was flooding the system with secondary material," therefore "degrading production, making recognition of significant information more difficult in the mass of the trivial" (Kerbel & Olcott, 2010). By 1976 the community was lamenting that this "information explosion" was "degrad[ing] the overall effectiveness of [intelligence], since there is simply too much to read, from too many sources" (Olcott, 2010). Yet, therein lays the critical enabling power of the automated software algorithms and computing systems that have evolved to handle the information filtering and aggregation tasks that a half-century ago were manual processes. Computational methods tend to perform more accurately, with greater statistical confidence in their findings, on larger datasets, so the very trend that was a limiting factor with human analysis has become an enabler of the big data revolution (Anderson, 2008).

3.4 MEDIA AS CULTURAL PROXY: ASSESSING REMOTE POPULATIONS

An emerging literature is validating the link between textually-derived latent indicators and human physiological states. For example a 2011 study in Science (Macy & Golder, 2011) demonstrated that the "tone" of Twitter matches the natural circadian rhythm, beginning in a positive mood and trending towards negativity through the course of the day, with seasonal and weekday/weekend differences.

Bollen (2011) found that the level of calm/anxious language in Twitter matches current investor anxiety accurately enough that it can reliably forecast future movements of the stock market three days out. Mentions of "cholera" on Twitter and in web-based news reports offered an early warning of the 2010 Cholera outbreak in Haiti a full two weeks ahead of official government public health alerts (Chunara, 2012). The discovery that web searches for flu-related terms increase dramatically 7-10 days earlier than the CDC's own Influenza Sentinel Provider Surveillance Network led to the creation of Google Flu Trends. Spikes in flu-related web searches have been found to be correlated with rises primarily of pediatric flu visits, with a far lower correlation for adult visits, suggesting demographic segmentation is present in such indicators (Dugas, 2012).

3.4.1 REGIONAL VARIATION

Regional variation in language use offers insight into cultural influences and has long been a focal point of linguists. In fact the oldest continually-funded initiative of the US National Endowment for the Humanities is the 60,000-entry Dictionary of American Regional English (DARE), a hand-compiled dictionary of localized language use (Wasley, 2012). Emerging work is demonstrating that even these linguistic enclaves can be computationally determined in realtime from online sources. A 2010 study of Twitter (Eisenstein et al, 2010) showed that geotagged tweets reflect regional differences that match known linguistic slang and topical emphases of those regions. A 2012 study (Floating Sheep, 2012) found that geotagged tweets using the words "church" and "beer" can similarly be used to map overarching cultural narratives. A key finding of their work was that "counties with high numbers of church tweets are surrounded by counties with similar patterns and…counties with many beer tweets are surrounded by like-tweeting counties," demonstrating spatial homogeneity in culture. The Obama

reelection campaign made extensive use of this class of geographically-centered cultural research to assess “voter’s hopes and fears” (Issenberg, 2012).

3.4.2 CULTURAL VARIATION

Cultural changes in time have reflected a variety of shifts in society from a recentering of culture on the self (Twenge, 2012) to a “dumbing down” of American political speech (Ostermeier, 2012). The shift over the past century towards narratives based around the self is especially poignant in that it suggests that the written word may be increasingly reflective of individual self and thus rich in the emotions and beliefs of greatest use in latent forecasting. This pattern appears to hold true even for books, suggesting that “traditional” forms of communication may be far more indicative of beliefs and behaviors than previously thought, and social media may not be the only source for information on the individual (Twenge, 2012). At least in the arena of American political speech, the complexity of discourse (words per sentence) and its intended audience (reading level score) appears to have decreased nearly linearly over the past 75 years (Ostermeier, 2012). While it is unclear whether this may have an impact on human perception of that speech, it does mean this content is likely to be more amenable to automated processing. The 2011 “Culturomics” article in Science (Michel et al, 2011) explored a range of topics using a temporal view of language change in digitized books, from censorship to changing standards of celebrity. In fact, the field of “Culturomics” was developed to use automated analysis of large datasets to remotely assess not geographically-inaccessible populations, but rather temporally-inaccessible populations. While a remote media assessment of southern Afghanistan households may be faster and easier to obtain than a door-to-door field survey, a historian studying the nineteenth century American South has no other alternative but to use media and other historical records to proxy views, since the subjects of interest are no longer living.

3.4.3 EMOTION AND INFORMATION PROCESSING

There is a growing literature in the advertising and marketing disciplines exploring the interplay of emotion with the processing of information. Deng & Poole (2010) found that a reader develops an immediate emotional response based on the visual appearance of information and that that emotional response follows him/her through the consumption and processing of that information. They further found that a reader's current emotional state when beginning the reading task impacts their emotional reaction to new information. This secondary effect is also found in the acceptance and rejection of new material that conflicts with an existing worldview. Chiang et al (2008) found that it is not the tone, source, or other characteristics that make a reader pay attention to a message, but rather whether that message was expected from that source. A Republican newspaper endorsing a Republican candidate will have little effect on voter decision making, but a Republican newspaper endorsing a Democratic candidate will have a measurable impact. Thus, sudden shifts in baseline narratives may yield substantial predictive information.

3.4.4 EMOTION WITHOUT NETWORK-CONTEXTUAL KNOWLEDGE

Of equal importance, the literature studying the diffusion of new information across the media sphere (such as breaking news about an emerging situation) has found there is no need to model the underlying network structure of a media system. Bandari et al (2012) found that statistical models could predict whether a news article would receive increased readership at up to 84% accuracy using only characteristic traits of that article. This is an important finding, in that much of the previous literature on forecasting media popularity has been based on using historical data to infer the hidden network structure of how past articles have cascaded across the set of available news outlets. Media outlets

tend to actively watch each other for story ideas and local and regional news outlets often to take their cues from larger national and international outlets. Studies such as Yang & Leskovec (2010) have focused on using past news flows to infer which outlets watch which outlets and reconstructing the attention network that connects those outlets, but this is extremely complex and connections likely vary continuously over time. Thus, Bandari et al's (2012) finding that similar accuracy may be achieved without this additional process is significant and has been replicated across both mainstream and social media platforms and across languages (Tatar et al, 2012; Ahmed et al, 2013).

3.5 CORE RESEARCH QUESTIONS

As discussed in the previous chapter, a growing literature from the economics and marketing disciplines has shown latent indicators to have predictive value in the constrained arena of economic behavior, yet the author's previous Culturomics 2.0 study (Leetaru, 2011) that inspired this dissertation, together with Chadefaux (2012), are the only current studies in the unclassified literature exploring the ability of emotional indicators to forecast future political stability. This dissertation is therefore designed to extend that approach from forecasting the extremely rare occurrence of whole-country collapse and militarized conflict to more generalizable changes in the physical unrest trajectory of a nation. Through doing so, this work will explore the degree to which these predecessor effects are present in broader and smaller-bore day-to-day human social behavior as opposed to being limited to nation-scale shocks. In addition, the codification of transition patterns from the visual observation method of Culturomics 2.0 and Chadefaux's study, towards statistical models capable of issuing daily alerts, would allow a more detailed assessment of the specific patterns and thresholds leading to future behavior. The following

three research questions are therefore constructed to be the focus of this dissertation to explore this phenomenon in detail:

- **RESEARCH QUESTION #1. What latent signatures precede physical societal-scale behavior and manifest themselves in the media in a measurable way?** The most basic question revolves around whether there are measurable latent signals such as specific classes of language that immediately precede societal-scale physical behavior and that can be operationalized to allow robust forecasting of that behavior. This has been extensively demonstrated in the artificial environment of the financial sector, and for whole-country collapse and conflict in the social world, but does it hold true for the day-to-day ebbs and flows of behavior of society as a whole? The models of conflict discussed in the previous chapter would suggest such signatures should exist, but this component of the research will explore whether they are manifest in the editorial environment of the mainstream news media and if the signals are strong enough to be robustly extracted through automated means without background knowledge of each location/time period. This question underlies the other dimensions of this dissertation in exploring whether latent dimensions remotely assessed from the media through automated means can adequately forecast future behavior.
- **RESEARCH QUESTION #2. Are signatures universal across geographies, or keyed to each location and culture?** If such latent signatures exist, are they universal, holding globally, or do they vary across geographies and cultures? If the latter, are there patterns that appear to constrain these localized patterns and that might offer clues into the underlying processes that prevent their generalizability to a global model? This has both theoretical and operational impacts. If signatures are found to be universal, this suggests the existence of fundamental linguistic markers that transcend socio-cultural boundaries and simplifies the use of these

findings in actionable behavioral forecasting environments. If, however, signatures are found to vary across cultures and geographies, this provides an empirical basis for future theoretical work into the specific journalistic and socio-cultural processes that cause such differing responses, but complicates the transition to actionable use in that supplementary processes would be required to maintain the model's socio-cultural baselines as cultures change. The current literature is divided, with studies showing both that latent measures are independent of culture (Crimson Hexagon, 2011) and that they are culturally and geographically dependent (SAS, 2011).

- **RESEARCH QUESTION #3. Are signatures universal across classes of physical behavior and intensity levels?** Do the same latent signatures forecast all classes of physical behavior, simply at different intensity levels, or do different measures forecast different types of behavior? Are there classes of event types that cannot be robustly forecasted through latent measures or that can be forecasted exceptionally well? Are there particular thresholds of intensity that physical behavior must exceed before it can be forecast through latent dimensions? This research question will explore the boundaries of such forecasting approaches, providing guidance on the environments in which they perform better or worse and offering empirical evidence for future study of the emotional-communicative-behavioral link.

CHAPTER 4: METHODOLOGY: QUANTIFYING RHETORIC AND REALITY

In order to empirically explore the three core research questions introduced in the previous chapter, they must be operationalized into a set of quantitative measures and methods. As the previous chapter introduced, the narrow timeframe and lack of geographic diversity in the existing literature has resulted in a conflicted patchwork of results. A key requirement therefore lies in accessing a cross-national longitudinal quantitative database of physical unrest, together with a collection of latent linguistic indicators measuring the discourse preceding those behaviors. The global scale and long time horizon required suggests that the operationalization must support computational inquiry, indicating the analytical methods must be computationally-driven. Thus this chapter will explore the reduction of the research questions into a set of addressable computational inquiries to enable a data-driven empirical exploration of the hypothesis that latent narrative indicators expressed in news coverage have predictive value in forecasting subsequent physical societal-scale behavior.

4.1 QUANTIFYING SOCIETY

In order to quantitatively study the link between latent narrative and physical behavior at societal scale, a dataset is needed that captures both behavior and its narrative undercurrents across multiple countries in different regions and ethnographic contexts and with a long enough timeframe to yield statistically significant findings. In particular, a diversity of event types and latent measures are needed to allow for sufficient segmentation across event types (since it is expected that some event classes may be easier to forecast than others) and variation across latent measures (since the existing literature has

suggested that predictive results are dependent on the specific linguistic dimension being measured). Most importantly, both the narrative and physical indicators must be quantitative in nature, offering a discrete analytical construct defining an occurrence such as a "riot" in terms of measurable quantities like its date, location, and situating factors (Schrodt & Yonamine, 2012).

While the underlying components of such a framework exist, to date no single study has combined them together. Radinsky & Horvitz (2013) cluster related news articles to form "storylines" and construct classifiers to forecast their occurrence, while Hunt (1997) and Chadefaux (2012) use hand-constructed databases of major militarized conflict. All three were forced to make use of coarse definitions of unrest and could only focus on physical conflict or other mortality events like disease outbreaks due to the lack of large cross-national databases of low-intensity events such as peaceful protests or positive actions such as peace accords or aid promises. The only studies incorporating higher resolution physical behavior data have focused exclusively on physically-based forecasting, leveraging world knowledge or temporal patterns in event occurrence (Schrodt & Yonamine). There is therefore a need for a generalized framework that allows exploration of the latent precursors to physical behavior on a global and historical scale.

Today's world of rapid electronic distribution and effective-zero cost of communication has led to a plethora of open information sources that could offer potential sources of both narrative and behavioral data (Olcott, 2012; Bean, 2011). However, despite the rise of new forms of communication like social media, the global reach, long temporal horizon, and well-studied construction mechanisms of the mainstream media make it an ideal data source through which to understand human societal behavior where a multi-decade time period is needed (Olcott, 2012). Rather than being an objective chronicle of society, news "is a commodity" which is "manufactured" to "induce anxiety within the minds of readers

and views" (Worrell, 2011, p. 40). Conflict naturally induces high levels of anxiety and thus forms one of the twelve fundamental tenants put forth in Galtung & Ruge's 1965 study of what makes a given story "newsworthy". This is beneficial in the study of societal behavior, in that it means that the news media pays special attention to conflict and that it also contextualizes that conflict in a way designed to maximize its connection to the local cultural narrative.

> The Press lives by advertising; advertising follows circulation, and circulation depends on excitement. "What sells a newspaper?" … The first answer is "war." War not only creates a supply of news but a demand for it. So deep-rooted is the fascination in war and all things appertaining to it that … a paper has only to be able to put up its placard "A Great Battle" for its sales to mount up. This is the key to the proclivity of the Press to aggravate public anxiety in moments of crises.
>
> (Lasswell, 1971, p. 192 in Worrell, 2011, p. 40)

In addition to reporting on realized conflict, the media is also the only cross-national source for understanding equally-important unrealized tensions, in which actors come increasingly close to the brink of conflict, but resolve their differences just short of physical conflict (Chadefaux, 2012). Such latent tension may preserver for years, decades, or even centuries, contributing to broader cultural narratives that can subsequently be reignited, leading to more rapid onset of conflict (Olcott, 2012; Bean, 2011; Roop, 1969). In fact, the formalization of Western Open Source Intelligence (OSINT) was based on this notion of the identification of media-based cultural indicators that offered potential precursor signals of conflict (Roop, 1969).

Hunt (1997) performed one of the more extensive recent unclassified quantitative studies of the link between rhetoric and war, operationalizing Karl Deutsch's (1957) theory of leading governmental priming indicators. Under this model, governments cannot initiate interstate war without first priming

their citizenry for the forthcoming casualties and cost. Governments are therefore expected to utilize their influence with the news media well before the intended onset of conflict to prime the population with beliefs about the other nation aligned with the need for conflict. Hunt experimentally tested this hypothesis by manually coding the emotional profile of editorials in government-aligned domestic newspapers as proxies for elite beliefs and found they offered measurable precursor indicators of future conflict. Indeed, conflict is rarely truly spontaneous: as noted earlier, even "sudden onset" conflicts are the result of a long buildup of latent frustration over time. Bertrand (2004) notes that in Indonesia "the scale of riots, demonstrations, and ethnic and religious conflicts displayed an array of grievances that had been little detected in previous years…what [had been] missed were the rising signs of 'nibbling' at the regime's fundamental structure…" (p. xii).

Radinsky & Horvitz (2013) used document clustering to identify "storylines" in New York Times coverage, which they defined as "a set of topically cohesive ordered segments of news that includes two or more declarative independent clauses about a single story." They constructed a classification model for each storyline and applied it to earlier news coverage to see if it would have forecasted the subsequent event. Given that this approach requires the creation of a classifier for each potential future event to be forecasted, the authors narrowed their scope to only attempting to forecast storylines known to occur in the future, which limited their ability to test the model's false positive rate.

At the same time, efforts as early as the 1970's developed similar processes for identifying and extracting codified records of political processes and physical unrest (Schrodt & Yonamine, 2012). So-called "event coding systems" use teams of humans or software programs to read large volumes of news media and compile quantitative lists of physical behavior, placing events into theoretically-informed taxonomies of societal action. A news report of a riot in Baghdad becomes an entry in a

spreadsheet recording the date, location, and actors involved and potentially connecting it to preceding and subsequent events. Modern computer-based coding systems can process tens of millions of articles in a matter of hours, recognizing several hundred types of events. The high resolution of these datasets, recording the specific actors and locations where events take place, and codifying even small non-violent actions, offers the potential for much finer-grained analyses compared with the large-bore militarized conflicts of previous work.

4.2 NEWS SOURCES

There are many large collections of news media content available today for computational analysis. For example, the International Conference on Weblogs and Social Media (ICWSM) 2011 Data Challenge collection "contains over 386M blog posts, news articles, classifieds, forum posts, and social media [posts]" covering January 13 to February 14, 2011, totaling more than 3TB of text (ICWSM 2011, online). However, despite their size, such collections typically span only a few weeks' of time – insufficient for establishing statistical significance in resulting temporal patterns. Online news aggregators like Google News collect hundreds of thousands or even millions of news articles per day from across the web. Google supplements this with a historical back-file that was leveraged in Chadefaux's 2012 examination of linguistic precursors to conflict in the twentieth century (Chadefaux, 2012). However, in his study Chadefaux specifically noted that he was limited in the analytic methods he was able to apply because of the inability to directly access the underlying text beyond simple keyword queries.

LexisNexis Academic Universe is one of the most widely-utilized news content aggregators in academic research, featuring almost 5,000 different news sources stretching back nearly four decades

("LexisNexis", online). Unlike online services like Google News, LexisNexis allows users to download the full text of each article for academic research, bypassing the limitations encountered by Chadefaux (2012). In particular, the ability to access the underlying full text of each article permits the application of sophisticated automated processing tools, including event coding systems, to the articles. Unsurprisingly, LexisNexis has therefore become the primary data source for quantitative event coding projects dating back several decades. While the service carries numerous newswires, among the most popular are Agence France Presse, Associated Press, Xinhua, Reuters, United Press International, and BBC Monitoring (Schrodt, 2010; Reeves, Shellman & Stewart, 2006). Due to contractual changes, Reuters is no longer available in LexisNexis and so is no longer widely used for event coding (Schrodt, 2010), while United Press International coverage of international events has decreased to a few hundred articles per month since the early 1990's. BBC Monitoring has become an increasingly-popular source for event coding (Reeves, Shellman & Stewart, 2006) due to its inclusion of translated local broadcast news from around the world. However, while Leetaru (2011) demonstrated significant conflict precursor signals in the service, for this dissertation it was desired to limit the analysis to services which product their own reporting, rather than aggregate existing local reporting. In this way, a more direct test of underlying narrative construction and how regional and cultural differences may be reflected in news reporting could be performed. Thus, in the end, LexisNexis Academic Universe was used to access the Agence France Presse, Associated Press, and Xinhua newswires, representing the largest news services in America, Europe, and Asia. It should be noted that while the next chapter relies on only the last decade of coverage from each source, it was not possible to make this determination until the full historical sequences had been processed to determine their geographic focus and event distributions.

4.2.1 AGENCE FRANCE PRESSE

Agence France Presse is one of the largest news agencies in the world and is the largest in France. It is also one of the primary sources used by Western intelligence services to monitor the continent of Africa (Leetaru, 2010). LexisNexis describes the newswire in its Source Information directory as follows:

> Agence France Presse is the world's oldest news agency. Based in France, with staffers and stringers in 129 countries, AFP offers a unique perspective on the world's news. AFP's Europe coverage is outstanding, its reporting from Africa is renowned and its Latin American correspondence comprehensive. AFP also covers the Middle East, Asia and the Pacific Rim.

LexisNexis coverage of the newswire does not begin until May 1991. An extensive manual review of the source indicated it did not include an overrepresentation of coverage of domestic French affairs, focusing instead primarily on international coverage. This suggested there was no need to incorporate additional filtering to specifically remove articles discussing French affairs. In addition, Agence France Presse articles occasionally quote French governmental officials on their views towards an emerging situation, which would result in many relevant articles being discarded if keyword-based filtering was used to remove all articles mentioning "France" or "French." While the newswire contains SECTION() metadata tags used to identify the major news desks such as sports and financial news, these are not always properly applied. In addition, major sporting or financial news is often treated as general news, rather than being tagged under the appropriate section heading. A manual review of a random selection of articles from each month was used to develop a lexicon of sports and financial-related keywords most commonly used in articles not properly tagged with the appropriate SECTION() tag. Thus, the following Boolean query was used to retrieve all articles from the "Agence France Presse – English" file in LexisNexis:

NOT section(sports) AND NOT section(financial) AND NOT golf AND NOT baseball AND NOT football AND NOT basketball AND NOT tennis AND NOT cycling AND NOT cricket AND NOT rugbyu AND NOT volleyball AND NOT "formula one" AND NOT subject(sports) AND NOT subject(financial results) AND NOT subject(economic news) AND NOT subject (stock indexes) AND NOT industry(stock indexes)

Figure 1 plots the total number of articles per month available in the LexisNexis archive of the newswire, showing that the newswire underwent steady growth through a peak in March 2001 and steadily decreased its output over the subsequent decade through mid-2010. It has largely remained constant at an average of around 8,000 articles a month over the last three years. There are also several outages visible in the first two years of its appearance in LexisNexis, which is noted on its LexisNexis Source Information page. In all, LexisNexis records 2,135,896 articles totaling 661,009,337 words through September 2012, averaging around 309 words per article.

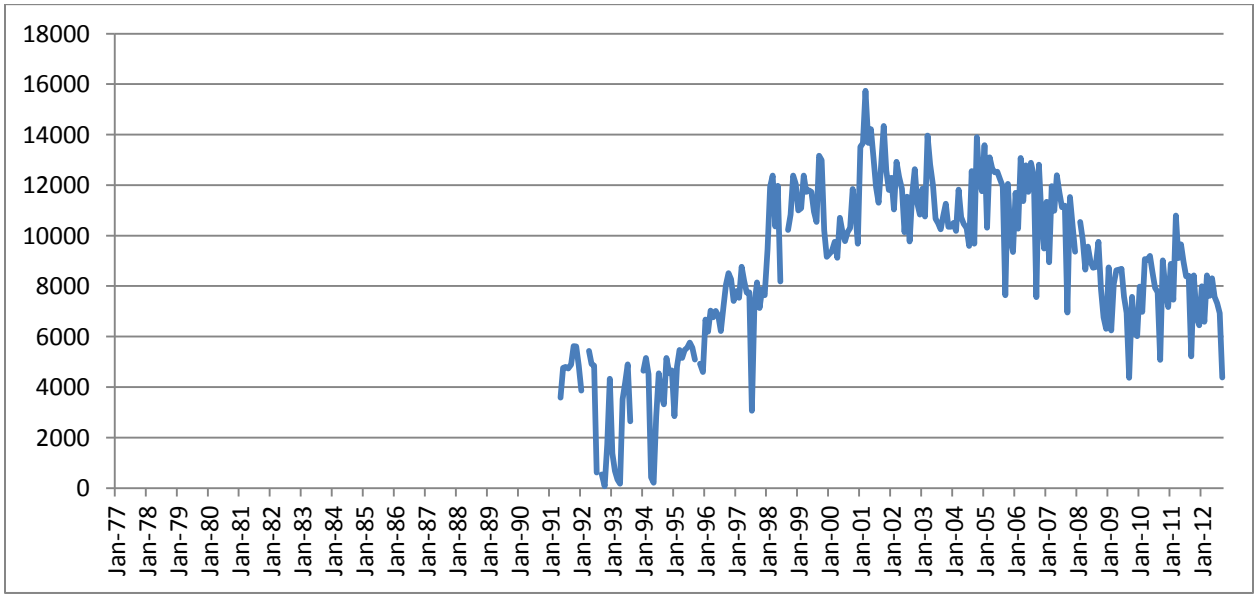

Figure 1 - Articles per month published by Agence France Presse

Table 1 shows the top 25 countries most frequently discussed by Agence France Presse, ordered by the percent of all articles published by the service through September 2012 that mentioned each country.

Here, any mention of the country or any city or other geographic landmark within the country was counted. Overall, there is a clear emphasis by Agence France Presse on Europe and the Middle East, with a particular focus on the United States and Great Britain.

Table 1 - Top 25 countries by percent of all Agence France Presse articles mentioning that country

Country	% All Articles
united_states	10.44
united_kingdom	4.79
france	4.31
russia	3.56
china	3.41
israel	3.10
iraq	3.08
germany	2.72
japan	2.36
india	1.78
iran	1.73
afghanistan	1.73
pakistan	1.64
italy	1.57
egypt	1.33
australia	1.31
indonesia	1.25
spain	1.22
turkey	1.21
south_korea	1.15
belgium	1.15
syria	1.07
canada	1.03
saudi_arabia	0.98
lebanon	0.98

4.2.2 ASSOCIATED PRESS

The Associated Press is one of the largest news agencies in the world, operating 243 bureaus across the world. Unlike Agence France Presse, the Associated Press is operated as a cooperative, in which any story published by a member news agency is automatically redistributed and available for any other member to publish. LexisNexis describes the newswire in its Source Information directory as follows:

> Founded in 1848, and now delivering news and photos in over 100 countries, The Associated Press sees itself as the oldest and largest news service in the world. The AP is a nonprofit cooperative (i.e., a member-owned organization) with its roots in the newspaper industry. Regular members of The AP are obligated to report exclusively to The AP news that breaks locally, but might be of interest to the media elsewhere in the U.S. or overseas. This system gives The AP a news gathering reach well beyond what would be possible with only its staff resources. Coverage includes international news, national news (other than Washington-datelined stories), Washington news (only stories of national interest), business news, and sports.

The Associated Press newswire contains a wide assortment of news that heavily emphasizes domestic United States events, including local and regional newspaper coverage. Beginning in December 1978 the newswire added SECTION() metadata tags that allow filtering of coverage to just national or international stories (prior to this date there was no choice but to download all coverage). The newswire also has a special designation of “top news” used to identify major breaking or important stories regardless of their geographic focus. Thus, the following Boolean query was used to retrieve Associated Press coverage from LexisNexis:

> *"top news" or section(international)*

Figure 2 plots the total number of articles per month available in the LexisNexis archive of the newswire, reflecting the lower volume of coverage compared with Agence France Presse. The sharp drop in coverage volume between November and December 1978 reflects the introduction of the new SECTION() metadata tag that allowed retrieving just international articles. While the newswire underwent steady growth during the late 1990's, it has experienced a decade-long decline in its international coverage, stabilizing at around 1,600 articles per month over the last three years. In all, LexisNexis records 944,483 articles totaling 358,398,400 words through September 2012, averaging around 379 words per article. Table 2 shows the top 25 countries in terms of the relative percentage of all Associated Press coverage during this time that mentioned each. As with Agence France Presse, there is a significant emphasis towards European and Middle Eastern countries and a similar emphasis on French coverage.

Those familiar with the Associated Press will likely question why the primary Associated Press newswire was used here, rather than the specialty Associated Press Worldstream newswire, which is exclusively focused on international news coverage. LexisNexis does in fact offer an archive of the Worldstream service that begins in October 1993 that is comparable in terms of daily coverage volume to Agence France Presse. However, for unknown reasons the LexisNexis archive of Worldstream ends abruptly in July 2010, with coverage past this date exclusively carrying sporting results.

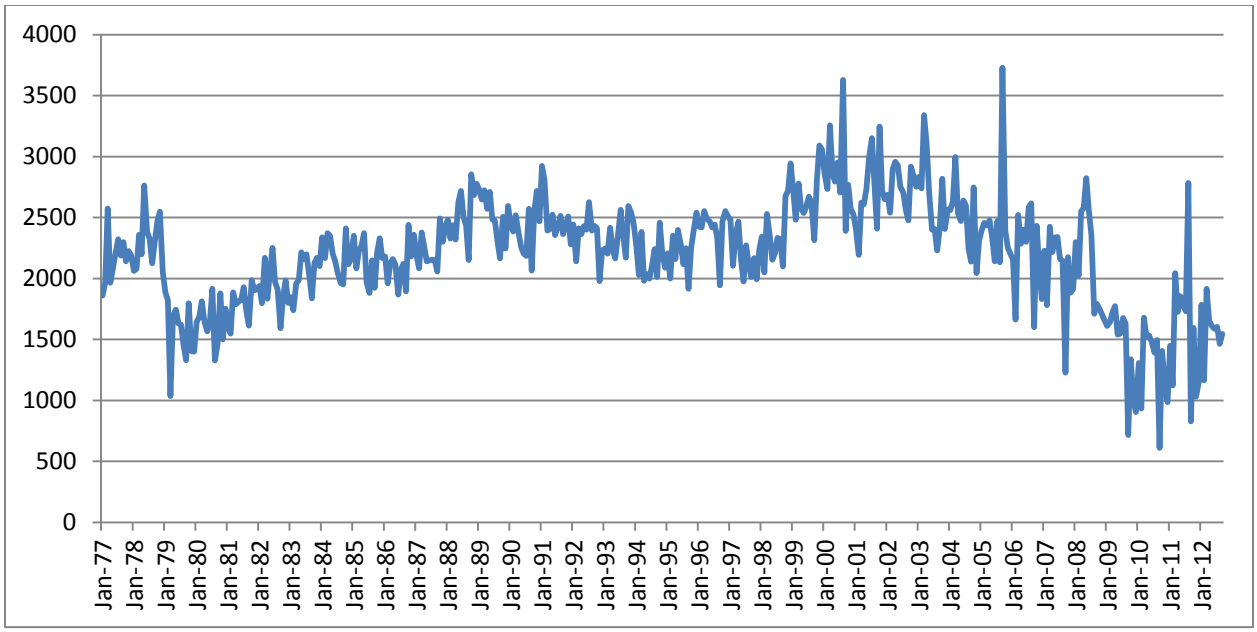

Figure 2 - Articles per month published by Associated Press

Table 2 - Top 25 countries by percent of all Associated Press articles mentioning that country

Country	% All Articles
united_states	13.35
united_kingdom	5.32
russia	4.51
france	3.39
israel	3.36
germany	2.74
iraq	2.71
china	2.27
japan	1.94
iran	1.87
italy	1.87
egypt	1.50
afghanistan	1.44
lebanon	1.44
canada	1.27
pakistan	1.26
india	1.26
spain	1.15
syria	1.14
west_bank	1.12
saudi_arabia	1.05

Table 2 (cont.)

Country	% All Articles
mexico	1.04
south_africa	0.98
poland	0.97
turkey	0.97

4.2.3 XINHUA

Xinhua is the official news agency of the People's Republic of China and the largest news service in the country, operating 107 bureaus around the world. While it still retains its official role in promulgating the views and statements of the Communist Party, XInhua has considerably expanded since its founding in the 1931 towards a general-purpose global news service competing with services like Reuters (Troianovski, 2010). LexisNexis describes the newswire in its Source Information directory as follows:

> Xinhua is the authoritative source for information on Chinese government affairs, economic performance and Chinese views on world affairs. All Western news correspondents in Beijing rely on Xinhua's English-language news report to keep abreast of Chinese affairs. The agency reports on Chinese affairs, including the economy, industry, trade, agriculture, sports and culture. Coverage includes diplomatic changes and extensive international reporting often from Africa or the Middle East. Xinhua also provides useful coverage of non-Chinese Asia.

As its description above suggests, Xinhua has an extensive focus on domestic Chinese news, which would overemphasize China over other countries. Through manual review of a random selection of articles from each month, it was determined that adding in exclusion keywords to drop those articles mentioning either "China" or "Chinese" removed domestic coverage with a minimal false positive rate. Unlike Agence France Presse, Xinhua coverage of international events uses quotes from Chinese officials more sparingly, meaning this filtering criterion has a minimal impact on international coverage. Xinhua also has a dedicated financial newswire called Xinhua Economic News Service, separating Xinhua's

extensive coverage of the financial markets from the Xinhua General News Service newswire used here. It does not, however, offer the SECTION() tags used with Agence France Presse and Associated Press coverage to filter out sports-related coverage, necessitating the use of additional keyword filters. Thus, the following Boolean query was used to retrieve Xinhua coverage from LexisNexis:

> *NOT china AND NOT Chinese AND NOT olympic AND NOT snooker AND NOT boxing AND NOT hockey AND NOT marathon AND NOT motorcycling AND NOT soccer AND NOT handball AND NOT cycling AND NOT tennis AND NOT world cup AND NOT basketball AND NOT wrestling match AND NOT wrestling score AND NOT iceskating*

Figure 3 plots the total number of articles per month available in the LexisNexis archive of the newswire. The near-tripling of coverage between December 1998 and November 1999 reflects the US involvement in Iraq during this period, which attracted a singularly large volume of coverage from Xinhua. The service also has two major outage periods in LexisNexis, from April 1995 to June 1996 (inclusive) and April 2008 to October 2008 (inclusive), so those are removed from consideration for all analyses. In all, LexisNexis records 1,699,442 articles totaling 332,043,292 words through September 2012, averaging around 195 words per article. Table 3 shows the top 25 countries in terms of the relative percentage of all Associated Press coverage during this time that mentioned each. As with Agence France Presse and the Associated Press, there is a strong emphasis towards European and Middle Eastern countries.

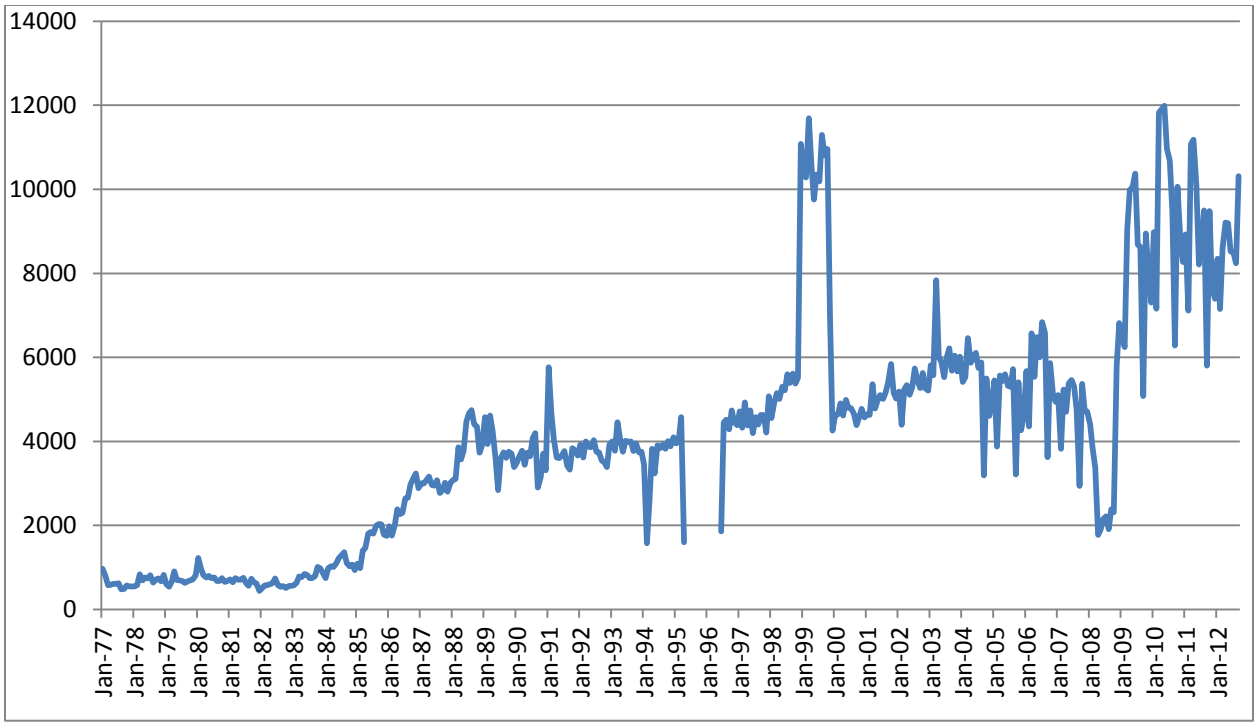

Figure 3 - Articles per month published by Xinhua

Table 3 – Top 25 countries by percent of all Xinhua articles mentioning that country

Country	% All Articles
united_states	10.76
israel	3.81
united_kingdom	3.34
russia	3.27
iraq	2.96
france	2.43
japan	2.32
pakistan	2.01
india	1.98
egypt	1.94
iran	1.94
germany	1.92
afghanistan	1.88
thailand	1.53
philippines	1.44
south_africa	1.42
australia	1.41
turkey	1.37
indonesia	1.34

Table 3 (cont.)

Country	% All Articles
syria	1.26
lebanon	1.14
nigeria	1.08
west_bank	1.07
kenya	1.07
italy	1.06

4.2.4 COMPARING THE SOURCES

Throughout this dissertation the three news sources will be analyzed individually in order to tease apart their mutual differences and explore whether predictive features found in one source are universal across the others. In total across the three sources, 4,779,821 articles were processed in the course of this dissertation, totaling 1,351,451,029 words. Figure 4 shows the Z-scored (standard deviations from mean) plots for all three sources overlapping their relative growth and decay patterns. Figure 5 shows the combined monthly article volume across the three sources, demonstrating in particular the significant mutual growth during the 1990's. The three sources, however, are poorly correlated in their temporal profiles. Even restricting the analysis to only overlapping periods of time, the monthly coverage volume of Agence France Presse has a Pearson correlation of r=0.27 with Xinhua and r=0.35 with Associated Press, while Xinhua and the Associated Press are correlated at r=0.03. While weak, Agence France Presse is correlated with the other two sources at $p < 0.0005$, indicating statistical significance, with Agence France Presse and the Associated Press being the only two sources not to have a statistically meaningful correlation. In terms of geographic emphasis, the three sources are closely aligned, with Agence France Presse and the Associated Press being correlated at r=0.97 in terms of the relative percentage of each's coverage dedicated to each country, while Agence France Presse is correlated at r=0.94 with Xinhua. Xinhua and the Associated Press are correlated at r=0.95. All three

are therefore at the highest level of statistical significance ($p < 0.0005$), indicating there are no significant differences in their respective geographic focus.

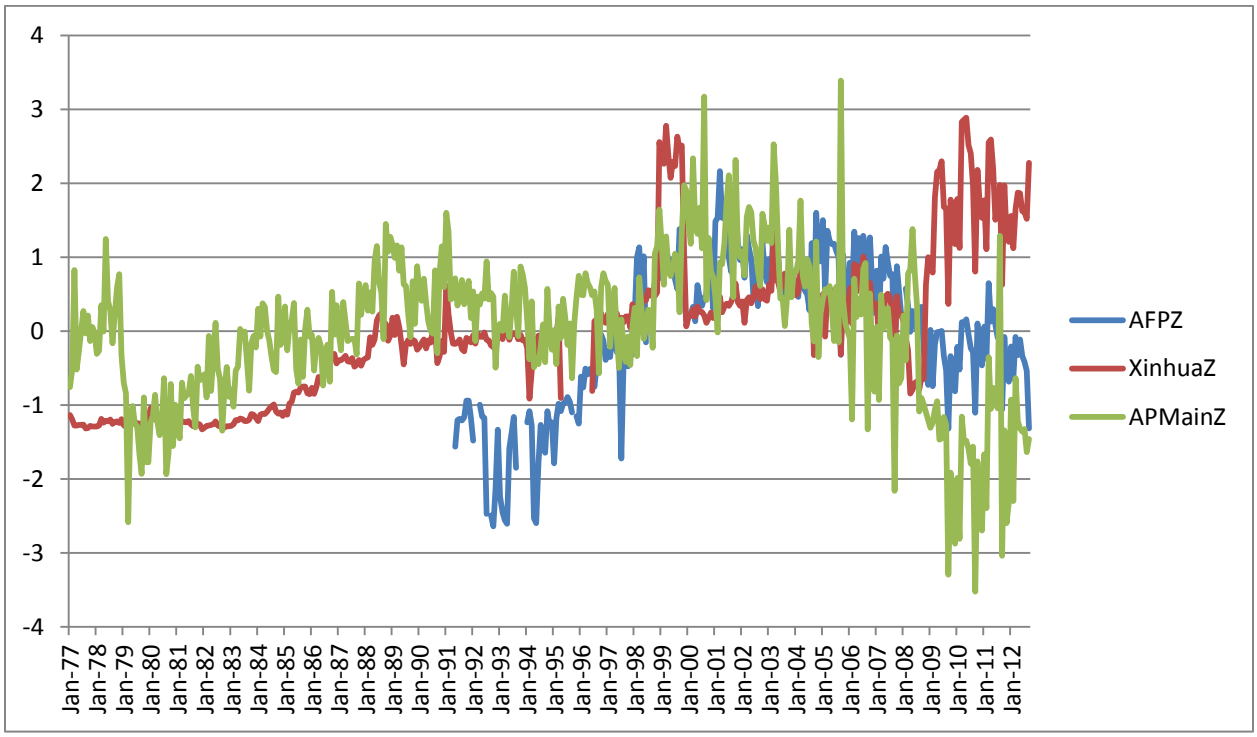

Figure 4 – Z-scored articles per month comparing Agence France Presse, Associated Press, and Xinhua

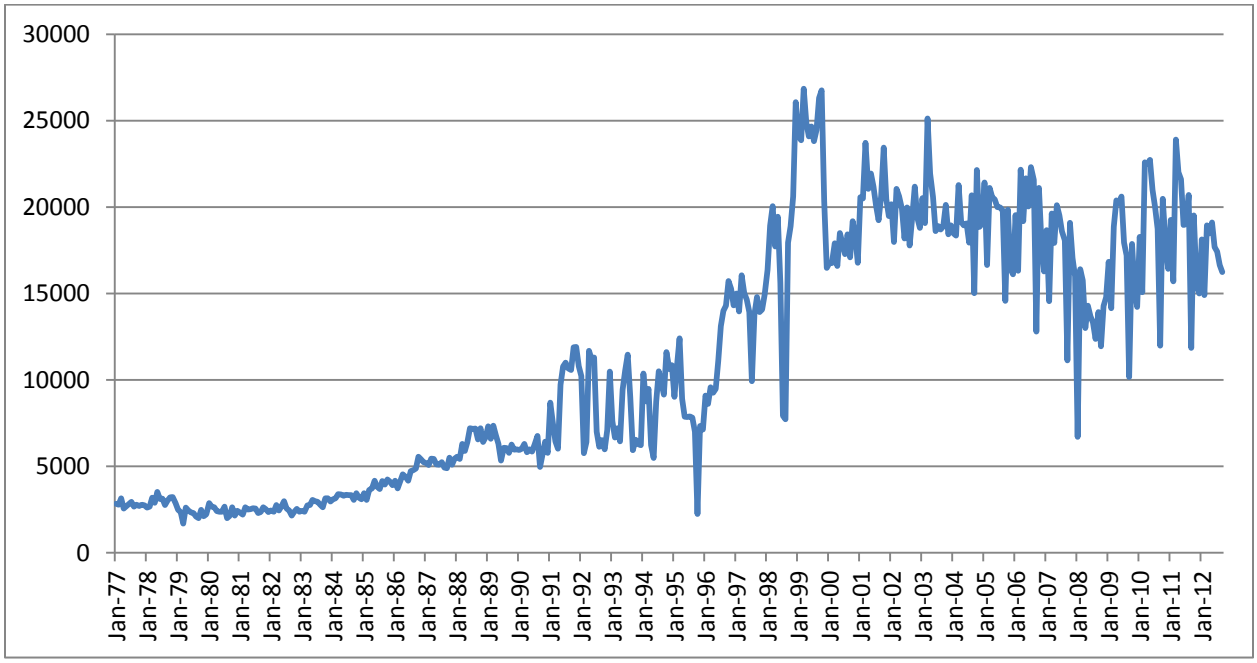

Figure 5 – Combined articles per month across Agence France Presse, Associated Press, and Xinhua

4.2.5 ADDITIONAL POST FILTERING

Despite careful construction of the queries used to retrieve each newswire from LexisNexis, including extensive manual review of random selections of content to develop exclusion keywords, a non-insignificant volume of sports and financial coverage was still retained. Such coverage can create complications for the event coding process in that it often contains language very similar to that used to describe violent political events, such as two sports teams "battling it out" or a company's stock price "under siege" (Schrodt & Yonamine, 2012). Thus, after all matching articles were downloaded from each newswire, a second manual review was performed across the combined content pool to filter any remaining sports or economic-related coverage. While this filter may eliminate certain economic-related articles that might reflect or drive public opinion (such as an economic boom or bust), incorporating this filter dramatically reduced the number of irrelevant articles. Thus, the following Boolean query was applied in a post-processing stage after the content was downloaded, but before it was made available for secondary processing.

> *NOT boxing AND NOT hockey AND NOT marathon AND NOT motorcycling AND NOT soccer AND NOT handball AND NOT cycling AND NOT tennis AND NOT worldcup AND NOT world cup AND NOT basketball AND NOT wrestling match AND NOT wrestling score AND NOT icestaking AND NOT ice staking AND NOT skiing AND NOT football AND NOT coach AND NOT hockey AND NOT box office AND NOT snooker AND NOT cricket AND NOT game console AND NOT gaming console AND NOT tv show AND NOT bond market AND NOT currency trade AND NOT closed up AND NOT closed down AND NOT industrial average AND NOT nasdaq AND NOT dow jones AND NOT halftime AND NOT half time AND NOT the game AND NOT stocks declined AND NOT market declined AND NOT inflation AND NOT interest rate AND NOT regional growth AND NOT car sale AND NOT truck sale AND NOT midsize car AND NOT inflation AND NOT singer AND NOT teammate AND NOT team mate AND NOT freethrow AND NOT free throw AND NOT show times AND NOT athletic AND NOT free throw AND NOT touchdown AND NOT the season AND NOT rebounds AND NOT quarterback AND NOT point guard AND NOT fourth quarter AND NOT on the road AND NOT season high AND NOT diet AND NOT title bid AND NOT mixed doubles AND NOT bowl game AND NOT retail price AND NOT book review AND NOT garden AND NOT goalkeep AND NOT goal keep AND NOT mega million AND NOT megamillion AND NOT mega-million AND NOT lottery game AND NOT lottery winner AND NOT ticket sale AND NOT lottery jackpot AND NOT baseball AND NOT golf AND NOT growth outlook AND NOT the dollar AND NOT bank index*

AND NOT nfl AND NOT nhl AND NOT nba AND NOT sports AND NOT championship AND NOT entertainment

4.3 CODIFYING SOCIETAL BEHAVIOR

In order to computationally measure the link between latent narrative and subsequent physical behavior, a quantitative definition of behavior is required that resolves a physical act into a discrete record of occurrence. As Schrodt & Yonamine (2012) survey, the quantification of societal-scale behavior has its roots in the 1970's with the development of theoretically-informed taxonomies of political behavior such as diplomatic exchanges, protests, and violent conflict. The construction of such frameworks allowed the humanistic concept of a "protest" to be described in the quantitative terms of its attributes, such as location, date, and actors involved, codifying it as a discrete "event." While enabling statistical treatment of conflict for the first time, such taxonomies also paved the way for the modern computational modeling of conflict.

4.3.1 EXISTING EVENT DATABASES

The growing popularity of quantitative study of conflict has led to a growing array of so-called "event databases." The Armed Conflict Location & Event Dataset (ACLED) (Clionadh et al, 2010) offers more than 50,000 georeferenced events covering more than a decade for an array of countries in Africa, Asia, and the Middle East. However, it offers only a handful of event categories, primarily arrayed around military action such as troop movements and clashes. More specialized collections include the Global Terrorism Database (GTD) from the University of Maryland (LaFree & Dugan, 2007), which contains more than 98,000 terrorist attacks around the world over the last three decades and the National

Counterterrorism Center's Worldwide Incidents Tracking System (WITS) database (Wigle, 2010), covering global terrorism over the last half-decade. While the two databases contain substantial coverage of worldwide terrorist activity, they are exclusively focused on such violence and most of their incidents occur in a small number of regions. Another possible dataset is the Peace Research Institute Oslo (PRIO) Comprehensive Study of Civil War (CSCW) project's Armed Conflict Dataset (Gleditsch et al, 2002). The Armed Conflict Dataset provides a longer time horizon than ACLED, GTD, or WITS, but includes only formal battlefield deaths and requires a threshold of 25 annual deaths for inclusion. More importantly, however, it does not include individual discrete event records, but rather aggregates conflict indicators into an annual binary indicator of whether the country was in conflict or not.

Alternatively, there are several turnkey monitoring services that include both news content and event records, such as the European Media Monitor (EMM) (Atkinson & Van der Goot, 2009) or DARPA's Integrated Conflict Early Warning System (ICEWS) (O'Brien, 2010). EMM monitors over 150,000 news articles from nearly 4,000 news websites each day, updating every 10 minutes. Its primary purpose is to automatically cluster articles about the same event together, allowing multiple perspectives on an emerging situation to be viewed. It provides a rudimentary event extraction service that identifies the primary event from each article, together with automatic binning of articles into a small set of predefined themes. However, similar to ACLED, it supports only a small set of coarse event types and thematic categories, primarily tied to European Union priority topics. It is also limited to just the most recent few years and does not permit downloading of the data (only the major trending records may be viewed).

DARPA's ICEWS project is perhaps the closest to the needs of this dissertation, as it makes use of an extensive taxonomy of more than 300 categories of events and is based on a rich theoretic framework

of human behavior that has been developed over several decades. Unfortunately, while the project has been described in unclassified open academic forms and limited subsets of its data have been used in academic publications, it is currently available only for official US military operational use. However, the core of the ICEWS system that performs the actual identification and codification of event records from ICEWS' open news feeds is a largely unmodified version of the open source TABARI software package from Pennsylvania State University, which will be explored in the following section.

4.3.2 TABARI AND CAMEO

One of the most widely-used event codification systems today is the TABARI software program (formerly known as KEDS) by Philip Schrodt at Pennsylvania State University. The TABARI system is an open-source software program that accepts a collection of news articles as input and automatically processes them using a grammar-based parsing system to identify more than 300 categories of societal behavior. For each event, it outputs a dyadic record recording that Actor1 performed a given action to Actor2, and the date and location of the action (Schrodt & Yonamine, 2012). In cases where multiple actors are involved, such as a multiparty peace summit, multiple entries are created, recording all of the dyadic connections. TABARI is a self-contained and fully autonomous coding system, operating in unattended batch mode, accepting as input a collection of news articles and outputting a tab-delimited file containing a list of extracted events.

TABARI has traditionally been applied only to the lead sentence of each news article, recording just the most "important" event in each story. However a growing body of literature has demonstrated the need for full-story coding to adequately cover many regions, in which the entire article text is coded and all events identified, and that approach is taken here (Schrodt, Simpson & Gerner, 2001; Huxtable,

1997). TABARI ordinarily records events at the level of the "country-day" in which more precise spatial and temporal information is discarded. This is problematic during periods of intense conflict, such as the 2011 Egyptian revolution when there were protests throughout the country. Thus, in keeping with Schrodt & Yonamine (2012), the fulltext geocoding process of Leetaru (2012) is used to associate each record with the closest geographic location in context. This preserves the geospatial resolution of high-intensity conflicts, ensuring they are properly recorded. Multiple mentions of the same event in the same or different articles are collapsed during a deduplication process to ensure that high-profile events attracting significant media attention are not counted multiple times.

While the news media traditionally reports on events that took place the previous or current day, they can on occasion report on future events or those in the significant past. TABARI has a dedicated date resolution system that automatically recognizes references like "two months ago" or "next week" and uses a calendaring system to resolve these to their appropriate date. To prevent forward-looking references from skewing the forecasting results (ie an article on a Monday stating that a peace summit will be held the following Friday), all events with forward offsets (ie occurring in the future) are excluded from the forecasts in the next chapter. Thus, only events occurring the day of the news article or in the past are included.

TABARI places its events into the Conflict And Mediation Event Observations (CAMEO) event taxonomy, which has its roots in a successive series of event taxonomies stretching back several decades and is one of the most widely-used taxonomies (Gerner et al, 2002). The latest version of the CAMEO event taxonomy used here (version 1.1b3) consists of 310 distinct event categories such as Code 1661 "Expel Or Withdraw Peacekeepers", Code 1832 "Carry Out Car Bombing", or Code 0311 "Express Intent to Cooperate Economically." These fall under the following 20 root categories:

- 01: Make Public Statement
- 02: Appeal
- 03: Express Intent to Cooperate
- 04: Consult
- 05: Engage in Diplomatic Cooperation
- 06: Engage in Material Cooperation
- 07: Provide Aid
- 08: Yield
- 09: Investigate
- 10: Demand
- 11: Disapprove
- 12: Reject
- 13: Threaten
- 14: Protest
- 15: Exhibit Force Posture
- 16: Reduce Relations
- 17: Coerce
- 18: Assault
- 19: Fight
- 20: Use Unconventional Mass Violence

Since many of these categories, especially Unconventional Mass Violence, will contain relatively few events for most countries, and to simplify the study of political dynamics, the CAMEO framework offers

the concept of "Quad Classes" (Gerner et al, 2002) that aggregate the individual categories to the "general behavioral level." Categories 01 through 05 are grouped under the heading of "Verbal Cooperation", categories 06 to 09 under "Material Cooperation", 10 to 14 under "Verbal Conflict", and 15 to 20 under "Material Conflict." This offers considerable simplification when exploring the kinds of broader trends of interest here.

A significant benefit of the TABARI system is that it allows the same news content to be used for both the analysis of narrative and the construction of the event database. This is important, as it ensures that any forecasting error is directly related to the model itself, rather than reflective of a disconnect between the narrative and event sources. Take the example of Agence France Presse news coverage of Egypt being used to forecast ACLED's Egyptian event records. If the resulting accuracy of the models was low, it would be difficult to determine whether the poor accuracy was because Agence France Presse coverage does not contain strong predictive narrative indicators, or whether it was because the ACLED database listed event types that the news agency did not cover in detail. Using the same news source for the entire processing pipeline avoids this potential source of error.

While any coding system, machine or human-based, will suffer from a certain level of error in codifying essentially qualitative occurrences into precise quantitative records, the combined TABARI + CAMEO system was the only system to pass the rigorous tests of the DARPA ICEWS competition (O'Brien, 2010). Under this competition, a wide array of state-of-the-art event recognition systems were applied to the same corpus of mainstream news articles and required to automatically recognize and codify all events within. The TABARI + CAMEO system performed so well that it is now the coding system used in the ICEWS United States Department of Defense operational watchboard, which compiles a daily list of political events worldwide to assist US military and intelligence analysts monitor global stability.

4.3.3 PROCESSING PIPELINE

This leads to the following processing pipeline used to construct the event database used here:

1. All relevant content from each newswire is downloaded from LexisNexis Academic Universe using the source-specific query defined earlier in this chapter.
2. Each article is subjected to the additional post-processing Boolean query to drop remaining sports- and financial-related news coverage.
3. Each article is subjected to fulltext geocoding from Leetaru (2012) to identify and disambiguate all geographic references contained in each article.
4. The TABARI system is applied to each article in full-story mode to extract all events contained anywhere in the article and the TABARI geocoding post-processing system is enabled to georeference each event back to the specific city or geographic landmark it is associated with.
5. The final list of events for each newswire is internally deduplicated. Multiple references to the same event across one or more articles from the same newswire are collapsed into a single event record. To allow the study of each newswire individually, events are not deduplicated across newswires (externally deduplicated).

4.3.4 EVENT DATABASE

In all, there were 28,877,172 events identified and coded by TABARI from the three news wires: 14,433,748 from Agence France Presse, 7,811,104 from the Associated Press, and 6,632,320 from Xinhua. On average, there were 7 events per article in Agence France Press, 8 from Associated Press, and 4 from Xinhua. However, since the three sources have different average article lengths, when

calculating the average words per event, the results are more even: 46 words per event on average for both Agence France Presse and the Associated Press and 50 words per event for Xinhua.

The three figures below show the total number of events per month for each news source in the four Quad Classes of Verbal Cooperation, Material Cooperation, Verbal Conflict, and Material Conflict. The gap in Associated Press events for 1992 is due to a technical error with the content downloaded for that year that prevented it from being properly processed, but since this is before the time period analyzed in the next chapter, it did not affect the results. Associated Press coverage, seen in Figure 7, has strong stratification among the four classes, making it easier to see their mutual patterns and that they are closely aligned. Indeed, the four series are correlated at between r=0.80 and r=0.97 ($p < 0.0005$) for all three news wires. Finally,

Table 4 shows the relative breakdown of all events into the four Quad Classes for each newswire. It is clear that Verbal Cooperation events are the most common, followed by Material Conflict, Verbal Conflict, and Material Cooperation. In all, across the three sources, 60.56% of events were Verbal Conflict, 17.34% were Material Conflict, 13.12% were Verbal Conflict, and 8.99% were Material Cooperation.

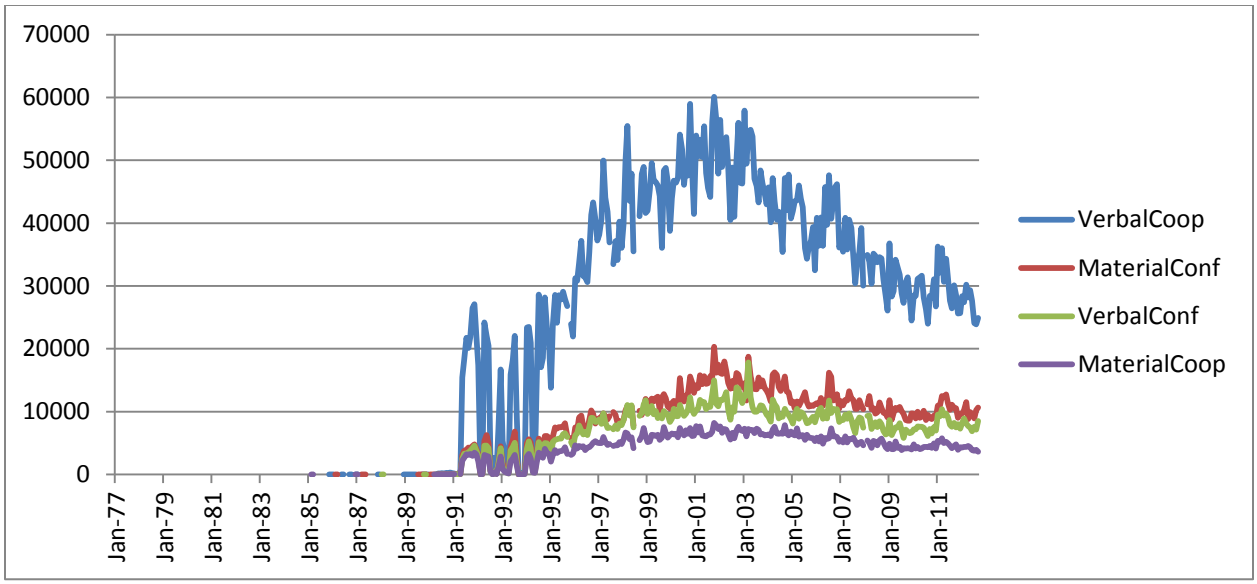

Figure 6 - Agence France Presse events per month by Quad Class

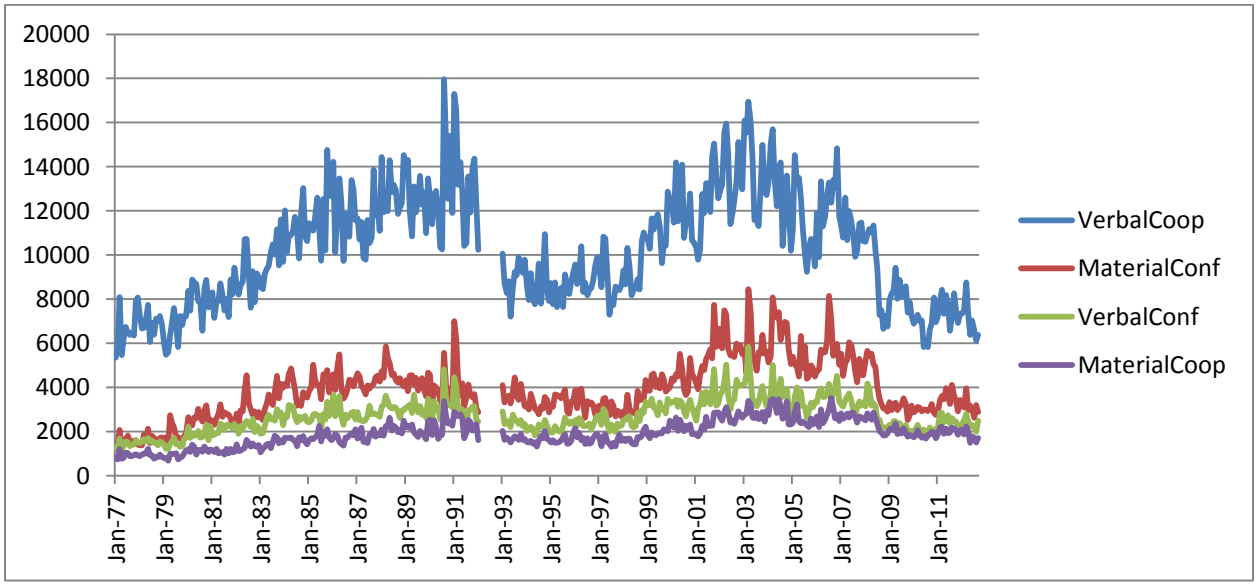

Figure 7 – Associated Press events per month by Quad Class

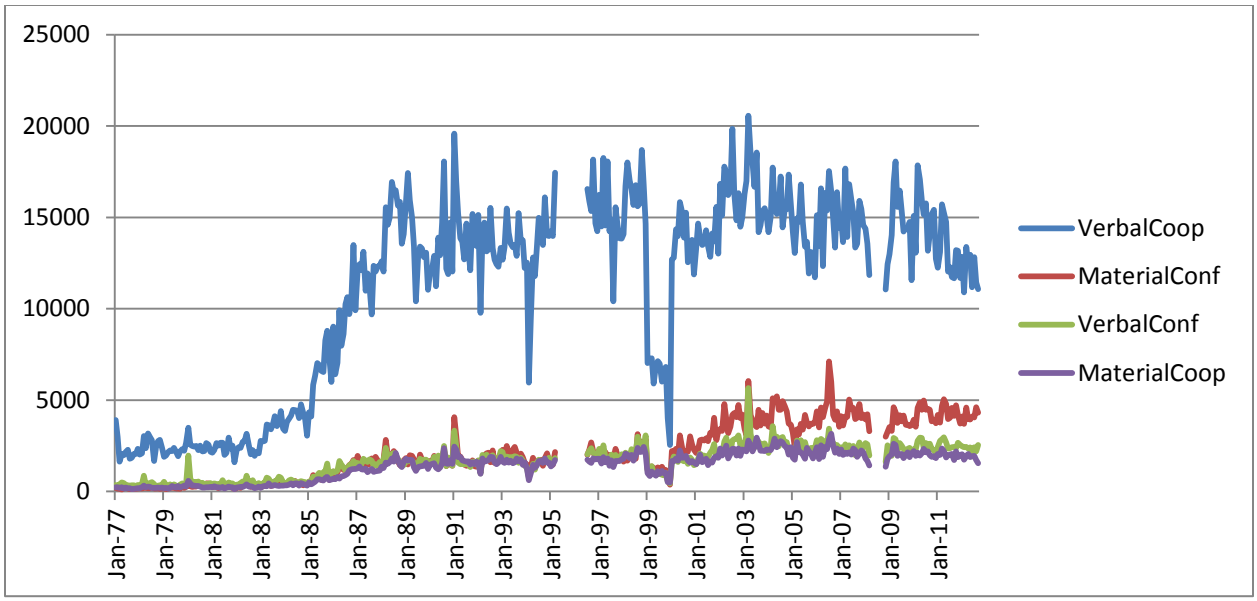

Figure 8 - Xinhua events per month by Quad Class

Table 4 – Breakdown by percent of all events in each newswire in each Quad Class

	AFP	AP	Xinhua
Verbal Cooperation	60.35	54.63	68.02
Material Conflict	17.37	20.75	13.22
Verbal Conflict	13.75	14.45	10.16
Material Cooperation	8.53	10.17	8.60

CHAPTER 5: FORECASTING BY CLASSIFICATION

One possible vision of a political forecasting system would require no theoretic model of societal unrest or human subject matter expertise: an analyst would simply select a small example list of previous incidents of interest and ask the system to forecast the risk on a day to day basis that similar incidents might occur again during specific time periods in the future. To do so, the system would observe a historical baseline of all available input streams preceding those selected past incidents, construct a set of machine learning models that most accurately retroactively forecast those previous events based on the data that was available at the time, and then apply those models to forecast future incidences of those events on a day-to-day basis, all without requiring any human assistance beyond selecting the incidents to forecast. By utilizing machine learning techniques to construct the model autonomously, rather than relying on human subject matter experts, the models can induce the natural underlying patterns of the data, potentially even evolving over time by comparing each day's forecast with actual unrest and adjusting the model accordingly.

The notion of providing a set of examples of a topic of interest to an algorithm and having it find new examples in the future is actually already widely used today in venues from personalized web portals to Google News (Antonellis, Bouras & Poulopoulos, 2006) in the field known as "text categorization" (Sebastiani, 2002). A typical configuration involves a user selecting a small set of example documents defining a category of interest (such as "basketball") and a small set of documents similar but not in the category (such as other sports-related material) as counter-examples. These two collections are converted to vector-space representations (Salton, Wong & Yang, 1975) in which a matrix is created where the rows are the documents and the columns are the set of all words appearing across those documents, with each cell in the matrix storing how many times a given word appeared in a given

document. These are then passed to a machine learning algorithm that identifies patterns in the usage of each word and how often documents with that word appear in the category of interest. For example, a typical categorization model, known as a "classifier," would likely determine that the words "rebounding" and "basketball" occur most often in basketball-related documents, while the words "score" and "audience" appear in both basketball and non-basketball documents with similar frequency, and "football" and "president" appear very infrequently in basketball documents.

To construct the actual classification model, a wide variety of algorithms are available, with popular ones including Support Vector Machines, Neural Networks, Random Forest, and Naïve Bayesian (Caruana & Niculescu-Mizil, 2006). While slowly being supplanted by more recent techniques like Random Forests, Naïve Bayesian classifiers remain among the most popular basic classifiers due to their relatively high accuracy on a wide array of text, fast execution speed, and simplicity of operation (Rish, 2001). In particular, all classification models have an array of parameters that can be adjusted to influence their operation, and the smaller set of parameters of Naïve Bayesian models makes them simpler to work with than other techniques like neural networks, while still achieving reasonable accuracy (Caruana & Niculescu-Mizil, 2006).

This suggests one possible approach to forecasting unrest could be to treat it as a classification problem, in which the "documents" are the text of news articles mentioning a specific country of interest on a given day and the categories to classify the documents into consist of various thresholds of the number and type of events that occur the following day. In this way, forecasting events in Egypt using Xinhua news coverage involves treating each day as a single "document" containing the text of all Xinhua news articles published that day mentioning Egypt in some way. The category of the document (day) is assigned to be the number of events that occur the following day, using a threshold to convert the

continuous event count variable into a binary variable of either a “High Event” or “Low Event” category. In training mode, a specific country or other area of interest, type of event and threshold for defining what constitutes a High Event day is selected by the analyst and the system uses available historical data (such as the previous several years) to construct its model. In forecasting mode, at midnight each day the system compiles all coverage of that country from a given news source and applies the model to generate a forecast of the likelihood that the following day will exhibit more than the specified threshold of events of interest.

Thus, rather than learning what words are most suggestive of basketball-related documents, the classifier learns what words are most suggestive of a certain number and class of events occurring the following day. Indeed, such a system generalizes and automates the process of hand selecting keywords used by Chadefaux (2012) to a fully automatic probabilistic approach that examines all words, using a lexical classification approach similar to Radinsky & Horvitz (2013), but arrayed around day-to-day ordinary events, rather than macro-scale occurrences. Since it requires no theoretic understanding of the underlying processes which might cause the events to be forecasted, relying instead purely on learned patterns of lexical probabilities preceding those events (Radinsky & Horvitz, 2013), such a classifier can be applied to any type of event, even ones for which little is understood about the underlying driving forces.

From an information-theoretic perspective, such classifiers essentially “learn” the surviving “information residue” or “information exhaust” left over from the underlying information environment in which knowledge about events both influences and is influenced by events. For example, at a peace summit, myriad political and economic factors weigh into each leader’s decisions as to which agreements to sign or not sign or whether to attend in the first place. While some of these factors may be readily

assessable, others may involve complex culturally-based calculations reflecting future knowledge or ambitions of that leader that resist traditional theoretically-based modeling (Ward, Greenhill & Bakke, 2010). Instead, lexical features capture the public information space surrounding those decisions, such as the way in which news media prime the public prior to major summits with large amounts of background information and expected outcomes, or reflect cultural narratives that envelope information streams (Olcott, 2012; Bean, 2011; Hunt, 1997; Deutsch, 1957).

5.1 MODEL CONSTRUCTION

Constructing the optimal classification model that yields the highest possible accuracy on a given collection for a given category requires considerable computation and is usually focused around the construction of a single or small number of models (Forman, 2003), whereas in this dissertation the goal is to construct reasonable models for a large number of different tasks. Complicating matters "because each parameter can affect performance, both singly and in combinations, many different [permutations] must be trained to adequately explore the parameter space" (Caruana & Niculescu-Mizil, 2006). In practice, this means constructing an optimal model can require sweeping across every possible permutation of all possible settings for a given algorithm. Thus, this chapter will focus itself not on exhaustively constructing the optimal model for a given scenario, but rather in using a coarser iteration metric to construct reasonably optimal models to reduce the computation time. Here the focus is not on constructing the best possible model for operational use, but rather on demonstrating the overall feasibility of classification-based forecasting, comparing different approaches, and understanding the underlying processes driving the variables that yield the greatest predictive power.

One might argue that a discriminative model like Logistic Regression would be preferred to a generative model like Naïve Bayes in the context here where the goal is not simply to construct reasonably accurate forecasting models, but to deconstruct their internal term lists to understand which terms are the most predictive of future unrest. In particular, it is clear that the Naïve Bayes assumption of term independence, in which each word appears with a probability that is entire independent of every other term (such that "obama" is assumed to appear with "president" no more often than it appears with "onion") is likely to be false in this context. The ability of discriminative models to decouple term interdependence would likely lead to higher raw accuracy and would be beneficial in the understanding of the language most predictive of various classes and prevalence of events. However, discriminative models do not always outperform generative models and can require substantially greater computational resources (Jordan, 2002). The size of the total term space and the number of models to be tested (over half a million configurations) necessitated the least computationally demanding solution and thus the Naïve Bayes approach is used here.

One of the more popular Naïve Bayesian implementations for basic classification work is the "e1071" implementation available in the R programming language, which makes use of the "SparseM" sparse matrix support library to minimize memory requirements (Dimitriadou et al, 2008). The "tm" library is commonly used with this implementation for construction and management of the term matrixes, including feature selection, addressed in more detail later (Feinerer, Hornik & Meyer, 2008). In this chapter, all text filtering is performed in PERL to leverage its robust pattern matching features, with the output of the filtering passed to R to use with "tm" and the "e1071" naivebayes methods. The use of R (Ihaka & Gentleman, 1996) for modeling allows the findings of this study to be expanded in the future to leverage the more than 4,000 contributed "packages" ("CRAN", 2013) representing most of the major machine learning algorithms and statistical methods in use today.

A Naïve Bayesian classifier approach to event forecasting has the following ten areas where it can be tuned or adjusted, each of which will be explored in this chapter.

- Text-Related
    - Number of previous days of text used to forecast subsequent events.
    - The type of text to use (full original text or filtered subsets of the text extracting words related to specific topics).
    - The amount of text to use (full text of article, lead paragraph).
- Event-Related
    - Number of future event days to forecast (forecast events for the following day, the following two days, following three days, following week).
    - Type of events to forecast (all events, Verbal Cooperation, Protests).
    - Threshold between High and Low Event days (how many events are required to count as a High Event day).
- Model-Related
    - Number of High Event training days.
    - Number of Low Event training days.
    - Term weighting.
    - Feature selection.

5.1.1 TEXTUAL SURROGATES

In addition to applying the classifier to the complete full text of each article, a cross-section of filtered sub-dimensions of the text are created in order to narrow the information environment and test how predictive each dimension is on its own. While a classifier will learn the relative probabilities of each term irrespective of the total number of terms, narrowing the available set of terms helps the classifier focus on those terms, where forecasts are based only on their respective probabilities, rather than muddying their predictive power in the noise of an array of unrelated words. There are thousands of open source dictionaries relating to emotions, themes, and other dimensions of text that may be relevant to conflict (IARPA, 2011; Leetaru, 2011; Chadefaux, 2012; Radinsky & Horvitz, 2013), yet the computational expense of constructing classifier models necessitates limiting the analysis here to a sampling of major categories of filters, rather than searching for a specific filter out of the thousands available that yields the highest possible forecasting accuracy. Thus, five major categories of indicators are used to construct filtered surrogates of the full text: emotion, part of speech, entities, event-based cues, and ethnic and religious group-based affiliation.

The original Culturomics 2.0 work (Leetaru, 2011) that inspired this dissertation explored the use of tone as an early warning indicator for impending country-scale governmental change. To test whether tone has similar applicability at forecasting smaller-scale events, three surrogates will be used that filter the news text down to only tone-related language. Traditional sentiment mining involves using scaled dictionaries to convert a collection of words to a single numeric indicator capturing its overall score along a specified emotional dimension (Whissell et al, 1986). Other than assigning words varying "intensities" such as capturing how "positive" or "negative" they are (Whissell et al, 1986), sentiment mining by its nature abstracts the resulting tonal score from the underlying text. For example, two

documents could both receive the same numeric "positive/negative" score, with one receiving the score for containing the words "rejected" and "expelled" and the other for "death" and "destruction." The former might be indicative of a failed diplomatic exchange, while the second might suggest a discussion of a natural disaster or military conflict. Simply graphing the overall tonal score over time as was done in Culturomics 2.0 renders this distinction invisible and makes it impossible to separate periods of negativity deriving from failed diplomatic exchanges from those of physical conflict. In addition, a text might contain both positive and negative extremes (such as "horrific" and "delightful" together in the same document) that combine to yield a neutral score, masking the high level of emotional content. Thus, nearly all tonal dictionaries offer the ability to compute the overall "positivity" and "negativity" of a text separately (Whissell et al, 1986).

Here the tonal dictionary from Shook et al (2012), measuring the density of "positive" and "negative" language in a text, is used to filter each document to create surrogates that contain only those words present in the dictionary. The first surrogate, FTXT_TONE, contains all words from the text that have entries in the tonal dictionary, essentially filtering the document for all "emotional" language. This version combines both positive and negative words. Two additional surrogates, FTXT_TONENEG and FTXT_TONEPOS contain only negative and only positive words, respectively, from the text, allowing a focus on whether "positive" (presumably relating to peace appeals and other diplomatic exchanges) or "negative" (likely resulting from conflict) language is most predictive of future behavior. Chadefaux (2012) found that specific hand-curated collections of negative language are highly predictive of future violent civil conflict, and these surrogates will enable the testing of this with emotional language more broadly.

Next, four versions of the text are created that filter for specific parts of speech. Part of speech tagging (Brill, 1992) involves applying algorithmic language models to annotate each word in a text with its appropriate part of speech, allowing decomposition of text along semantic roles. While most emotional words are adjectives, adverbs, or verbs (Whissell et al, 1986), filtering by part of speech permits a broader analysis of the role of descriptive language more generally in forecasting future unrest. Instead of looking only at adjectives predetermined to have specific emotional connotation, this allows all descriptive language to be isolated and explored on its own. Thus, FTXT_POSADJS contains just the list of adjectives and adverbs found in each document, complementing the emotional dimensions above by exploring whether it is emotion specifically or descriptive language in general that is more predictive.

Part of speech tagging also allows the removal of nouns from text, removing general names (all nouns) or person, organization, and other names (proper nouns). A 2012 study by the authors of the TABARI/CAMEO system applied Latent Dirichlet Allocation to a collection of news articles to determine how closely the CAMEO event taxonomy matched the underlying landscape of events actually described in the news media. They found that removing proper nouns significantly improved the results of their topic modeler by removing its reliance on specific people or locations that may be highly temporally-fixated (such as US Secretary of State Hillary Clinton being closely associated with certain countries during her tenure) (Bagozzi & Schrodt, 2012). This allows the classifier to focus on general discussion of leadership, rather than learning that a particular leader has become contentious. Two surrogates are therefore created that remove nouns: FTXT_POSNONOUNS removes all nouns from the text, eliminating any reliance on names or objects, while FTXT_POSNOPROPERNOUNS removes only proper nouns from the text. These test the findings of Bagozzi & Schrodt (2012) as applied to event forecasting. Finally, as detailed earlier, the TABARI system identifies events through a large lexicon of grammars that codify specific verb phrases into various event categories. TABARI captures only a fraction of English verbs in

its lexicons and many verbs not present in its dictionary may possess emotional connotations with potential applicability to forecasting, such as “wishing” or “loathing.” Thus, FTXT_POSVERBS contains just the list of all verbs (all verb tenses) found in the text. The Lingua::EN::Tagger PERL module (version 0.23) is used here for the part-of-speech tagging, applying a Hidden Markov Model-based engine.

In addition to part of speech tagging, several traditional natural language processing filters are explored, including removing “stop words” (common words that do not usually convey semantic meaning, such as “the”, “a”, “and”, etc) (Fox, 1989) and performing “stemming” (removing verb conjugations to convert a word like “running” back to “run”) (Lovins, 1968). Such filters are commonly applied to boost the accuracy of classification and information retrieval tasks, but can reduce accuracy if verb tense or certain stop words are more prevalent in some categories (Kantrowitz, Mohit & Mittal, 2000) (such as increased discussion of the future signaling an forthcoming peace summit). Both are tested here not with respect to any expected theoretic connection to conflict, but rather to determine whether they increase the accuracy of the classification models. The widely-used Snowball stemming engine and its 174-entry stop word list are used here (Porter, 2001), creating FTXT_STEMMED, FTXT_STOPWORDS, and FTXT_STEMMEDANDSTOPWORD surrogates.

Events in the CAMEO taxonomy revolve around people and organizations that perform actions or have actions performed upon them (the “actors”) and the places those actions occur (Schrodt, 2012). Intuitively it might seem that increases in news media attention to a particular political leader or terrorist organization could signify an important emerging situation, such as coverage of a terrorist bombing, hostage taking, or forthcoming political summit. Indeed, previous studies have found that specific political leaders tend to become closely associated with specific countries or regions over time (Bagozzi & Schrodt, 2012), suggesting the very mention of US Secretary of State Clinton with respect to a

nation is suggestive either of potential unrest there or of likely forthcoming diplomatic cooperation. To explore this further, the algorithm from Leetaru (2012) is used to extract a list of all person names, organization names, and disambiguated locations from each document and used to create META_NAME, META_ORG, and META_GEO surrogates. Since the R text classification system used here operates at the level of individual words, rather than phrases, all person and organization names are converted to single lowercase words by replacing all spaces and other non-letter characters with underscores, converting "Hosni Mubarak" to "hosni_mubarak" and "United Nations" to "united_nations." The META_GEO field is treated slightly differently. Here, both the name of the geographic landmark as it appears in the text and its standardized name from the United States National Geospatial-Intelligence Agency's GEOnet Names Server (GNS) and United States Geological Survey's Geographic Names Information System (GNIS) gazeteers are included (Leetaru, 2012). For example, a reference to "French soldiers" will result in META_GEO containing both "french" and "france." These three surrogates allow connections between the appearance of certain political leaders and organizations and the places they are affiliated with and future events to be explored in more detail.

While the Leetaru (2012) algorithms extract all person, organization, and place names, a final set of surrogates use TABARI's own ACTOR lists to extract only those names known to TABARI (and thus that could result in an event). TABARI will only identify an event in which both the verb phrase is recognized in its grammar lexicon and one of the actors involved in the verb phrase is found in its ACTOR list. This ACTOR list contains a list of recognized international and major national political leaders and organizations, but is far from exhaustive. Thus, the Leetaru (2012) algorithm will recognize more names than are present in TABARI's dictionary (in many cases these names are business leaders or average citizens that would not be expected to be found in the ACTOR dictionary). The FTXT_TABARIALLACTORS surrogate thus is a subset of the META_NAME, META_ORG, and META_GEO surrogates combined,

containing just those persons, organizations, and locations listed in TABARI's ACTOR dictionary that could form an event. Similarly, while the FTXT_POSVERBS surrogate mentioned earlier contains all verbs mentioned in the text, the FTXT_TABARIVERBS surrogate contains only those verb phrases contained in TABARI's VERB file. This narrows the list of verbs to only those that could possibly generate an event (ie, "wished" will be in FTXT_POSVERBS, but not FTXT_TABARIVERBS, while "attacked" will be in both). This tests the notion that increases in the use of event-related language (even if those specific mentions do not result in an event record) are suggestive of an environment in which future events may occur.

As a final test, a dimension that has a strong theoretic basis in instigating and enhancing conflict is explored. Ethnic and religious conflict is at the root of a substantial portion of modern societal unrest from Serbia to Sudan (Easterly, 2000). Indeed, the outsized role such conflict plays in national stability has led to the construction of a wide array of ethnic and religious "fractionalization indexes" that measure how culturally homogenous a given geographic area is and the presence of minority groups of certain size densities (Fearon, 2003; Alesina & Ferrara, 2004), which have become standard inputs in political and economic models of many regions. In addition, the author's previous Culturomics 2.0 work demonstrated strong upwards trending in ethnic group mentions ahead of ethnically-driven unrest (Leetaru, 2011). In fact, group-based tension has become such an important dimension in studying conflict that TABARI recently added new dictionaries that codify mentions of the world's major recognized religions and ethnicities (CAMEORCS and CAMEOECS, respectively) (Schrodt, 2012). The Version 1.1.b3 edition used here contains 1,392 name variants of recognized world religious groups and 614 name variants of world ethnic groups. Thus, the FTXT_RELIGIOUS and FTXT_ETHNIC surrogates contain only mentions of religions or ethnicities from the CAMEORCS and CAMEOECS dictionaries, narrowing the text to group-based discourse. Given the sometimes-contentious distinction between the

boundaries of ethnicity and religion (Todd & Ruane, 2009), a final surrogate, FTXT_ETHNICRELIGIOUS, combines the CAMEO religious and ethnic dictionaries together.

These five categories and the natural language processing filters result in 19 different surrogates for testing. Each day's text is passed through these filters to create these different subsets of the text to assess which offers the greatest predictive insight for a given country and news source:

- **FTXT_CLEAN**. This is the original raw text converted to lower-case with numbers and punctuation and other non-ASCII text removed. This tests the predictive power of the raw text itself with no additional filtering.
- **FTXT_ETHNIC**. This is the list of ethnic groups recognized by CAMEO. This tests the ability of ethnic-related language to predict unrest, acting as a proxy for ethnic tensions.
- **FTXT_ETHNICRELIGIOUS**. This combines the CAMEO religious and ethnic dictionaries to capture group-based discourse more broadly.
- **FTXT_POSADJS**. This uses part of speech tagging to compile the list of all adjectives and adverbs from the text.
- **FTXT_POSNONOUNS**. This uses part of speech tagging to remove all nouns from the text, including both proper and common nouns.
- **FTXT_POSNOPROPERNOUNS**. This uses part of speech tagging to remove all proper nouns from the text, while leaving common nouns.
- **FTXT_POSVERBS**. This uses part of speech tagging to compile the list of all verbs from the text.
- **FTXT_RELIGIOUS**. This is a list of all religious groups recognized by the CAMEO system.
- **FTXT_STEMMED**. This is the full raw text similar to FTXT_CLEAN, but where all words have been "stemmed" using the Snowball stemmer.

- **FTXT_STEMMEDANDSTOPWORD**. This is the same as FTXT_STEMMED, but with all "stop words" removed.
- **FTXT_STOPWORDS**. This is the same as FTXT_CLEAN, but with all stop words removed (and no stemming).
- **FTXT_TABARIALLACTORS**. This compiles all words found in the TABARI ACTORS dictionary.
- **FTXT_TABARIVERBS**. This compiles all of the words in the TABARI VERBS dictionary.
- **FTXT_TONE**. This compiles all words, positive or negative, found in the Shook et al (2012) tonal dictionary.
- **FTXT_TONENEG**. This narrows the focus of FTXT_TONE to consider only words in the negative list of the dictionary. This allows the model to focus on these words while excluding positive words.
- **FTXT_TONEPOS**. This is identical to FTXT_TONEPOS, but only includes positive words, while excluding negative words.
- **META_GEO**. This includes all locations from the text by applying fulltext geocoding from Leetaru (2012), identifying all locations from cities to local landmarks globally.
- **META_NAME**. This identifies all person names found in the text via Leetaru (2012).
- **META_ORG**. This identifies all organization names found in the text via Leetaru (2012).

It should be noted that the theoretically-informed process of selecting the indicators used here would at first appear to be in conflict with the anti-theoretic rhetoric of latent indicators. In particular, as discussed earlier, one of the arguable benefits of latent indicators is that they reduce the need for a detailed theoretic understanding of a class of behavior before it can be forecasted. Instead, several theories of conflict drivers and natural language processing have been used to select the dimensions evaluated in this dissertation. In actuality, a production forecasting task based on such latent indicators

would likely collect and evaluate every available data stream: potentially thousands or tens of thousands of available indicators. Indeed, the primary goal of the IARPA Open Source Indicators (2011) program is precisely to conduct such a broad-sweeping evaluation of all available unclassified latent indicators, including both theoretically-suggested and non-theoretically-suggested data streams. This dissertation explores a smaller number of such indicators and focuses on a mixture of theoretically suggested (group- and actor-based conflict) and linguistic (part of speech tagging) streams to constrain both data collection and computational requirements within the scope of a doctoral dissertation, and also to enable a more thorough analysis of the resulting models. However, it should be noted that an operational forecasting project would likely simply acquire every available data stream and every available filter and test each of them to determine which yielded the highest accuracy (IARPA, 2011).

5.1.2 FEATURE SELECTION AND WEIGHTING

All text classification systems begin by transforming their input text into a term matrix of words and documents. However, since even a word which appears just once in the entire collection becomes a term entry, the resulting matrix is both massively large and highly "sparse," meaning the majority of terms appear in only a few documents. Such sparse terms do not convey significant meaning in that a word appearing only once, on a High Event day, should not result in a 100% probability that documents containing that word indicate High Event days. Removing such terms both reduces the computational requirements of the modeling process and prevents them from biasing the resulting model. This process is known as "feature selection." In addition, each word must be assigned a "weight" indicating its "importance" to the collection. This is often the raw number of times the word appears or the total number of documents mentioning it, but can also incorporate the word's prevalence in the overall collection and other factors. Feature selection and weighting are significant factors influencing the

accuracy of a model and an area of significant ongoing research, usually customized for any given application (Yang & Pedersen, 1997; Forman, 2003).

The table below shows the average accuracy of two major weighting and selection approaches across the countries, sources, and configurations tested in this chapter. The TFIDF measures refer to calculating the Term Frequency Inverse Document Frequency (TF-IDF) weight for each word, which normalizes the total number of times a word appears by the total number of documents it appears in. This is especially helpful for source-specific stop words like "Xinhua" or "news agency," which are not listed in traditional stop word lists, but in the context of a given news source are equivalent to "noise" words. Here, the mean, first, and third quartiles of the TFIDF scores for all words were calculated and only those words with higher scores were retained as valid features (Grun & Hornik, 2011). The other three measures simply dropped those words which had less than 30%, 70%, or 99% sparsity, which means they appeared in at least 70%, 30%, or 1% of all documents, respectively. In addition, three major methods of term weighting were tested: raw term frequency, TFIDF weighting, and Chi-squared weighting. Chi-squared weighting did not yield substantially better results than TFIDF, while being more computationally expensive, while TFIDF on average was 1.9% more accurate than raw term frequency. In practice, the method which yielded the most accurate models was to weight all terms by their TFIDF scores for the Naïve Bayesian model, but to perform feature selection by a simple measure of term sparsity. All words which appeared in more than five documents (determined experimentally and as used in papers like Griffiths & Steyvers, 2004) or appeared in more than 1% of all documents (whichever was lower) were retained. This typically reduced the number of words retained for processing by 60-70% or more.

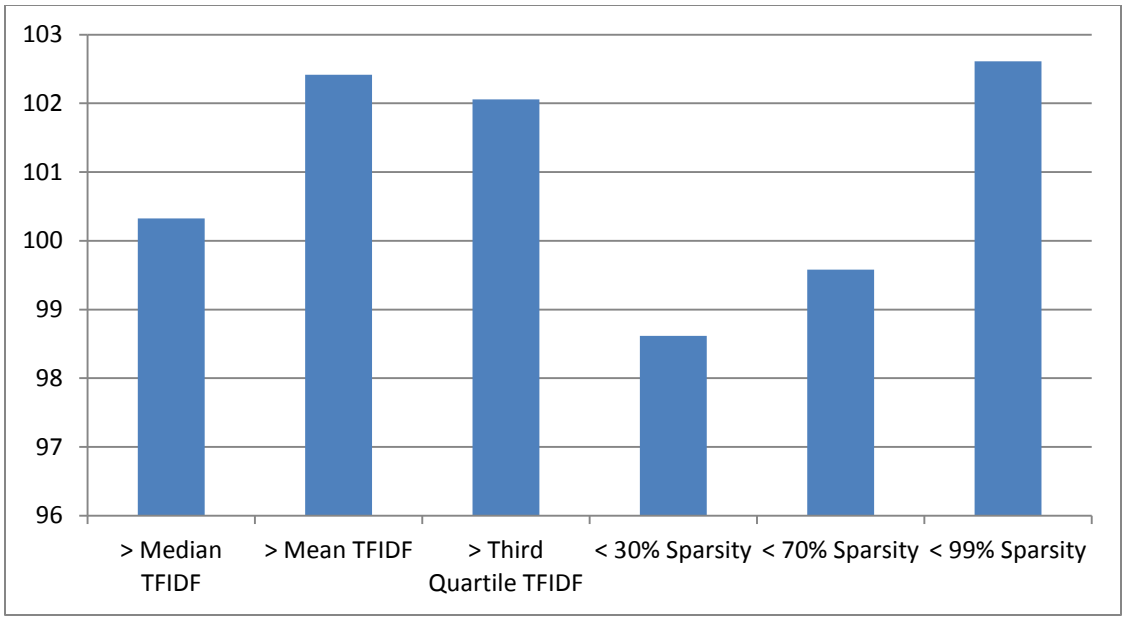

Figure 9 - Average accuracy of different feature weighting and selection approaches

Finally, as noted earlier, all text classification models are sensitive to the balance between topical categories: in this case the relative ratio between High and Low Event days during the training period (Montgomery, Hollenbach & Ward, 2012). Thus, for each model configuration tested, the system determines the total number of High Event and Low Event days during the training period and iterates over a set of permutations of those days, starting with, for example, 100 High Event and 100 Low Event days, then trying 100 High and 200 Low, then 100 High, 300 Low and so on, then repeating for 200 High and 100 Low, 300 High and 100 Low, and so on. This allows the system to determine the most accurate balance of training days.

Traditionally, N-fold validation is applied to a model in which the source data is broken into N chunks and alternating chunks are used for training and testing. However, this assumes the data is invariant with respect to time, whereas in this case the "documents" are actually days arrayed linearly in time, with more recent documents actually containing the mentions of previous events that were recognized and extracted by the TABARI system, resulting in an invalid test. In other words, if a given set of articles

were used to assess accuracy at forecasting the events from an earlier week, it is highly possible that those articles may have been the source articles that TABARI identified the prior week's events from. This would merely test the ability of the classifiers to replicate TABARI's event extraction system, not the ability to forecast future events.

5.1.3 ASSESSING ACCURACY

In order to compare different model configurations, a robust accuracy metric is required. The most basic measure of the accuracy of a classification model is the total number of correct forecasts divided by the total number of forecasts. This measure, however, assumes there is a relative balance between the categories being forecasted. If, for example, out of 1,000 days being forecasted, 990 were Low Event days, with just 10 High Event days, and an algorithm simply forecasted every day to be Low Event, this would result in in an Accuracy Score of 990/1000 = 99%. In one set of tests, the most accurate model had an 80.6% Accuracy rate, but upon further investigation, the model had simply predicted every day to be Low Event, and out of the 1,826 days in the test period, 355 were actually High Event. While this is technically correct, it strongly favors models which bias their error towards the dominant category (Weng & Poon, 2008). In practice, this invariably favored baseline classifiers which simply predicted every day to be High or Low Event, depending on the relative breakdown of which dominated the testing period.

More commonly, the output of a classifier is represented as a so-called Confusion Matrix containing four values: True Positive (High Event days correct forecasted as High Event days), True Negative (Low Event days correctly forecasted as Low Event days), False Positive (Low Event days incorrectly forecasted as High Event days) and False Negative (High Event days incorrectly forecasted as Low Event days). These

values are combined into Precision (the percent of High Event forecasts that were actually correct), Recall (the percent of High Event days that were actually forecast as such), also known as the True Positive Rate, and optionally the True Negative Rate (the percent of Low Event days correctly forecast as such). One of the most popular accuracy metrics used for assessing classifier accuracy involves calculating the harmonic mean of the Precision and Recall scores, yielding what is known as the "F1 score" (Rijsbergen, 1979). However, this too becomes significantly skewed towards the dominate category in cases where there is a significant imbalance between High and Low Event days (Yang & Liu, 1999). Demonstrating this problem, one test configuration involved forecasting 2,205 days, of which 2,190 were High Event days and 15 were Low Event days. The model with the highest F1 score correctly forecasted all 2190 High Event Days, but incorrectly classified 5 of the 15 Low Event days as High Event. Indeed, the models with the highest F1 scores tended towards test periods that were dominated by either High or Low days. Other methods, such as Area under the Curve (AUC) scores suffer similarly (Bradley, 1997). Indeed, construction and assessment of machine learning algorithms applied to highly imbalanced datasets is an entire field of study in itself (Maloof, 2003) and no single standard accuracy metric has yet emerged in the field of political forecasting for this reason (Brandt, Freeman & Schrodt, 2011).

One of the reasons that traditional metrics such as the F1 score perform poorly in this situation is that in traditional classification tasks, such as identifying documents by topic from a larger collection, the goal is often to achieve a balance between precision and recall (hence the use of the harmonic mean in calculating F1). However, when forecasting political events, it may not be desirable to balance the two. Biasing towards correctly identifying major events while missing smaller events may be important in some situations, while a high false positive rate that minimizes the risk of missing an event may be critical in others. There are a variety of variants of the F1 score that can bias away from F1's equal

balance towards either Precision or Recall, but this requires making a decision whether to penalize false High Event days or false Low Event days (Yang & Liu, 1999). In essence, one must decide whether it is better to have many false positives and not miss a major event (which risks the "crying wolf" syndrome in which an analyst fatigues of the stream of false positives and begins ignoring the forecasts) or to have few false positives, but miss major events (as happened with the Arab Spring). The specific circumstances of a given usage scenario will likely dictate the precise balance required, so it was desired not to have an accuracy metric that specifically penalized one over the other.

To best balance these competing needs and to address the extreme imbalance between High and Low event days in many test periods, a simple synthetic accuracy metric is used here, which sums the True Positive and True Negative rates together. The higher the score, the better the overall performance, with scores greater than 100% indicating better-than-random performance, and a score of 200% indicating perfect accuracy. This metric has the added benefit of equally penalizing False Positives and False Negatives in direct proportion to their density in the testing period.

Finally, not all countries will be covered by a given news source every day, or coverage may lack specific features, such as language representing a specific emotion, on certain days. A classifier cannot make a forecast in the absence of input, yet the absence of text also does not necessarily indicate the absence of events, as censorship, media fatigue, or other media effects can impact coverage volume (Reeves, Shellman & Stewart, 2006). A human analyst reading the news each day to make forecasts about the future monitors multiple sources, switching among them based on which has coverage on a given day, and does not make a forecast for a given area of interest in the absence of available information (Olcott, 2012). Thus, a similar approach is used here, where during both testing and training periods, days in which no text is available are dropped from consideration. In production use, such a system would

simply display a warning message on such days alerting the user that no forecast could be made for the area of interest from the given source.

5.2 FIRST EXPERIMENT: USING XINHUA TO FORECAST EGYPT

As the first test of the feasibility of classification-based forecasting, Xinhua was selected as the smallest of the three newswires, minimizing the computational cost of experimentation. Egypt was selected as the first country to study, as the previous Culturomics 2.0 work demonstrated strong emotional cues in its mainstream media coverage that suggest lexical cues may offer insight into future activity (Leetaru, 2011). It is also the 10th most-discussed country in Xinhua during this period, with 72,482 articles, and is the most mentioned Middle Eastern nation outside of Israel.

Xinhua's coverage of Egypt does not exceed more than a few articles per month until 1999, so the training period was set to January 1, 1999 through December 31, 2005, with the test period running from January 1, 2006 through December 31, 2011. This offers an even split of six years of training data and six years of testing data. For each day, all articles mentioning Xinhua that day were compiled and filtered into the 19 textual surrogates introduced earlier, and the day assigned into the High Event category if the following day had one or more events of any kind and the Low Event category if the following day had no events. To begin with, only the lead paragraph of each article was examined, as this reflects the most important portion of the story per the "inverted pyramid" style of writing used in professional journalism (Bell, 1991). Next, the total number of High and Low Event days were tallied for each type of text (ie, dropping days that contained no person names for META_NAME, for example) and models constructed with the various permutations on the number of High and Low training days

provided to the model, with the most accurate configuration reported in the table below. In this way, such a classifier, if used for operational forecasts, would be run each evening at midnight on the current day's news coverage to generate a forecast as to whether the following day would yield one or more events.

Table 5 - Egypt/Xinhua: single day of text predicting single day of events (1/1/1999-12/31/2005 training period, 1/1/2006-12/31/2011 test period)

Text Type	True Pos	True Neg	Accuracy	Pos + Neg
FTXT_TABARIVERBS	27.74	83.66	36.88	111.40
META_GEO	48.78	62.60	51.04	111.39
FTXT_POSNONOUNS	35.42	74.24	41.76	109.66
FTXT_TONENEG	32.88	76.45	40.01	109.34
FTXT_CLEAN	30.18	78.67	38.10	108.85
FTXT_STEMMEDANDSTOPWORD	28.72	80.06	37.10	108.77
FTXT_TABARIALLACTORS	48.51	60.11	50.41	108.62
FTXT_TONE	39.32	69.25	44.21	108.57
FTXT_POSVERBS	35.86	72.30	41.81	108.16
FTXT_STOPWORDS	30.18	77.84	37.96	108.02
FTXT_POSNOPROPERNOUNS	32.83	75.07	39.73	107.90
FTXT_STEMMED	28.29	79.22	36.61	107.51
FTXT_TONEPOS	33.89	73.33	40.34	107.23
FTXT_POSADJS	27.53	79.22	35.97	106.75
META_NAME	10.29	94.94	24.04	105.24
FTXT_ETHNIC	44.94	58.89	47.18	103.84
FTXT_RELIGIOUS	89.18	14.20	77.39	103.39
FTXT_ETHNICRELIGIOUS	39.72	62.82	43.43	102.55
META_ORG	0.00	100.00	16.27	100.00

A purely random classifier, which assigns each day in the test period at random to either a High or Low Event day achieves a perfect balance of 50% True Positive and 50% True Negative, leading to a combined Pos + Neg score of 100.00. Overall, it is clear there is a significant range in accuracy from META_ORG which simply assigns all documents to Low Event days (and thus achieves accuracy no better than random chance) through FTXT_TABARIVERBS, which achieves an 11.4% increase in accuracy over

random chance. As expected, there is an inverse relationship between the True Positive and True Negative rates. The FTXT_RELIGIOUS surrogate demonstrates why the traditional Accuracy measure is insufficient here, as it biases strongly towards False Positives and since there are more High Event than Low Event days here, this results in a biased accuracy score. Thus, the synthetic accuracy metric adopted here of adding the True Positive and True Negative rates yields a more reasonable assessment of accuracy. In addition, while 11% better than random chance might at first seem inconsequential, a truly random classifier cannot bias its error towards minimizing False Positives or False Negatives in line with the need of a given application. From the table above, it is clear that specific text dimensions allow one to bias in either direction as needed.

Of particular interest, the fifth most accurate classifier relies on the raw full text of the articles and is just 2.55% less accurate than the best model. This suggests that even a basic classifier relying on the full text, without using the theoretical knowledge about conflict or language used to construct the other textual filters, still yields an 8.85% increase over random chance. It is also interesting that removing stop words and performing stemming actually decreases the accuracy slightly in this case. The use of TDIDF term weighting likely accounts for the minimal impact of stop word removal, as those will automatically be assigned a minimal weight, while the slight negative impact of stemming likely reflects that certain verb tenses may indicate a greater emphasis on future action. Egypt has not had a recent history of significant ethnically or religiously-driven strife during the period analyzed here. While there certainly has been sometimes-violent tension between the State and the Copts, those tensions have been largely small-bore and localized (Fahim & Stack, 2011), with the majority of national unrest being secular in nature driven by economic or democratic interests (Kirkpatrick, 2011), and this is reflected in the low predictive power of those dimensions.

Returning to the discussion earlier regarding the ratio of High to Low Event days used as the training input, the table below shows the accuracy of the model trained for each combination of the number of High and Low training days, illustrating the dramatic effect on accuracy this ratio can have. It is clear that this ratio can be used to bias the model towards High or Low Event days to nearly any degree required. Thus, if minimizing False Positives is more critical than minimizing False Negatives, this can be achieved. The non-linearity of the training ratio demonstrates the criticalness of the parameter sweep approach to test all ratios of High/Low input documents when training the models.

Table 6 - Egypt/Xinhua: accuracy results of FTXT_TABARIVERBS training ratios of High/Low event days

Low Train	High Train	High/Low Train Ratio	True Pos	True Neg	Pos + Neg
100	100	1.00	4.54	95.57	100.11
100	350	3.50	10.22	91.14	101.36
100	600	6.00	18.88	86.98	105.86
100	850	8.50	15.74	88.37	104.10
100	1100	11.00	27.74	83.66	111.40
100	1350	13.50	26.34	83.93	110.27
100	1600	16.00	35.10	73.68	108.78
100	1850	18.50	41.32	63.71	105.03
100	2054	20.54	47.54	59.83	107.37
350	100	0.29	85.18	8.86	94.05
350	350	1.00	68.52	27.98	96.50
350	600	1.71	40.62	62.33	102.94
350	850	2.43	43.05	64.82	107.87
350	1100	3.14	45.00	61.22	106.22
350	1350	3.86	43.16	62.33	105.49
350	1600	4.57	50.84	55.40	106.24
350	1850	5.29	51.00	54.29	105.29
350	2054	5.87	55.22	50.14	105.36
477	100	0.21	73.23	23.55	96.77
477	350	0.73	71.17	26.04	97.21
477	600	1.26	44.40	61.77	106.18
477	850	1.78	47.21	62.05	109.26
477	1100	2.31	51.33	55.96	107.28
477	1350	2.83	47.32	59.56	106.88

Table 6 (cont.)

Low Train	High Train	High/Low Train Ratio	True Pos	True Neg	Pos + Neg
477	1600	3.35	50.73	55.68	106.41
477	1850	3.88	55.22	50.69	105.91
477	2054	4.31	54.68	51.25	105.92

The tables above use only the lead paragraph from each article to focus on the most "important" substance of each story. This could have an especial impact on the ethnic and religious categories, as such information is not likely to appear in the lead paragraph unless group affiliation plays a primary role in the events therein (such as an attack on members of a specific religion). The table below therefore repeats the process of above, but uses the full text of each article, instead of only its lead paragraph. For FTXT_CLEAN (the unaltered document text) this increases the size of the model input from 37MB to 119MB, quadruples the memory consumption, and more than doubles the execution time of each model from 20s to 45s on average. While the peak accuracy of the top model is increased by 0.47%, the average accuracy across all models is substantially decreased and most models are significantly less accurate than with lead text. As expected, ethnic and religious dictionaries do see a slight increase in accuracy, but the reduction in accuracy across other categories reinforces that the body of a news article largely contains supporting details secondary to the narrative of the lead paragraph (Bell, 1991). Thus, all subsequent analyses in this chapter will rely only on lead paragraphs.

Table 7 - Egypt/Xinhua: accuracy of full text rather than lead paragraph text

Text Type	True Pos	True Neg	Accuracy	Pos + Neg
META_GEO	53.98	57.89	54.62	111.87
FTXT_TONE	39.16	70.64	44.30	109.79
FTXT_TONENEG	40.85	68.70	45.40	109.54
FTXT_POSVERBS	33.53	75.62	40.41	109.15
FTXT_TABARIVERBS	34.07	74.79	40.72	108.86
FTXT_STOPWORDS	28.99	78.95	37.15	107.94
FTXT_TONEPOS	33.95	73.33	40.38	107.28

Table 7 (cont.)

Text Type	True Pos	True Neg	Accuracy	Pos + Neg
FTXT_CLEAN	29.15	78.12	37.15	107.27
FTXT_POSNOPROPERNOUNS	30.88	76.18	38.28	107.06
META_NAME	14.76	91.29	27.19	106.05
FTXT_STEMMED	37.05	68.98	42.26	106.02
FTXT_STEMMEDANDSTOPWORD	60.52	45.43	58.05	105.95
FTXT_POSADJS	31.91	73.96	38.78	105.87
FTXT_TABARIALLACTORS	29.91	75.62	37.38	105.53
FTXT_POSNONOUNS	30.72	74.79	37.92	105.51
META_ORG	11.50	92.35	24.65	103.85
FTXT_ETHNIC	44.94	58.89	47.18	103.84
FTXT_RELIGIOUS	89.18	14.20	77.39	103.39
FTXT_ETHNICRELIGIOUS	39.72	62.82	43.43	102.55

5.2.1 FORECASTING HIGH-EVENT DAYS

The tables above forecast only whether the following day will contain zero events or whether it will contain one or more events. The majority of the conflict literature such as Shrodt (2000) or Montgomery, Hollenbach & Ward (2012) instead attempt to forecast time intervals with "large" numbers of events, rather than attempting to forecast every day having at least one event. Thus, the table below repeats the analysis of Table 5, but this time, in addition to constructing models to forecast any day with an event, it also constructs alternative models that instead forecast days with larger numbers of events (adjusting the High/Low event threshold). It calculates the mean, median, and first and third quartiles of the number of events per day in the training period and tests each of those as the threshold for High Event days. This significantly improves the accuracy for all text types, with the highest accuracy being nearly 17% greater than random chance. It also achieves significantly better balance between True Positive and True Negative rates for many text types.

Intuitively, it makes sense that it should be easier to forecast days with larger numbers of events than those with fewer events, especially since an isolated protest occurring on a day by itself may have far less preceding media coverage than a massive surge of protests across a country, which are more likely to stem from an issue that has received greater media attention in the lead-up to unrest.

Table 8 - Egypt/Xinhua: Forecasting days with multiple events

Text Type	High Thres	True Pos	True Neg	Accuracy	Pos + Neg
FTXT_TABARIALLACTORS	16	51.04	65.55	62.40	116.59
FTXT_POSVERBS	16	46.88	69.31	64.43	116.18
FTXT_POSNONOUNS	16	47.92	68.21	63.80	116.12
FTXT_POSNOPROPERNOUNS	16	48.33	67.51	63.35	115.85
FTXT_TABARIVERBS	16	44.79	70.12	64.62	114.91
FTXT_STEMMED	16	52.50	61.97	59.91	114.47
FTXT_CLEAN	16	50.63	62.66	60.05	113.28
FTXT_STOPWORDS	16	49.17	63.82	60.63	112.98
FTXT_TONE	16	38.33	73.70	66.02	112.03
FTXT_POSADJS	16	42.92	69.02	63.35	111.93
FTXT_ETHNIC	16	28.78	82.80	70.98	111.58
FTXT_TONENEG	16	38.54	72.84	65.38	111.38
META_GEO	1	48.73	62.60	51.00	111.33
FTXT_STEMMEDANDSTOPWORD	16	48.33	62.49	59.41	110.82
FTXT_ETHNICRELIGIOUS	16	30.85	78.90	68.46	109.75
FTXT_TONEPOS	16	31.94	77.68	67.74	109.62
META_NAME	18	17.30	92.27	78.83	109.58
FTXT_RELIGIOUS	18	14.09	89.32	74.53	103.41
META_ORG	1	0.00	100.00	16.27	100.00

The forecasts above represent the worse-case scenario of forecasting an event to a specific day: if a riot occurs two days later instead of the following day, it is counted as an incorrect forecast. Few studies impose such stringent requirements on their forecasting models, instead allowing periods of a month or more to elapse while still considering the forecast correct (Radinsky & Horvitz, 2013). Thus, the graph below relaxes this threshold and redefines the forecasting criteria to forecast whether a specific number

events will occur the following day or within the following two days, three days, four days, or five days (again, testing the multiple High Event day thresholds from above). As expected, relaxing the window yields a significant boost in accuracy, up to 32.27% better than random chance when using META_GEO to forecast whether there will be one or more events in the following five days. Expanding the forecasting window to forecasting two to three days in advance yields only a minor improvement, with the greatest improvement coming at the four to five day window. This appears to be driven by the fact that four days already represents a substantial fraction of a week and thus begins to cluster isolated multiday sequences of events into single contiguous runs of events for the purposes of the accuracy metric.

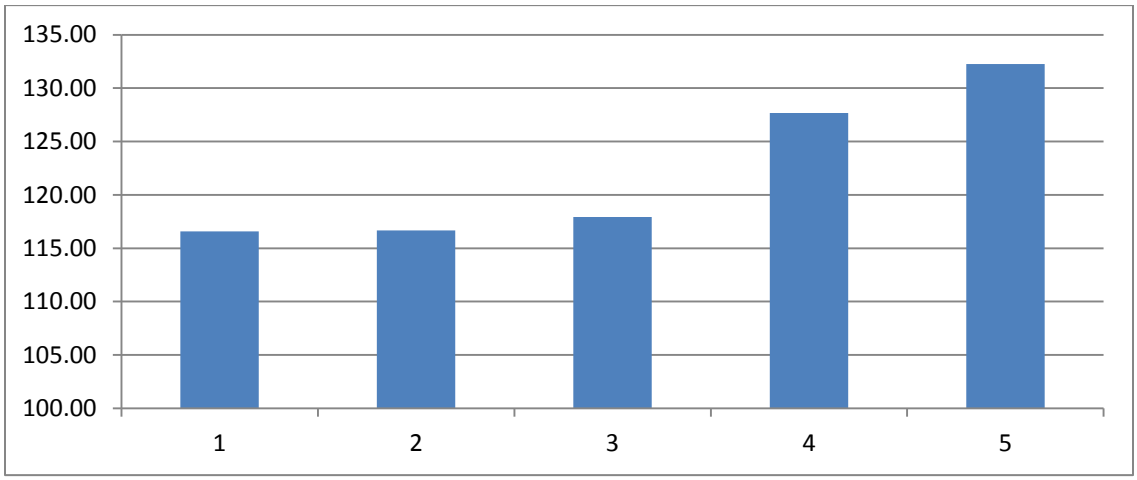

Figure 10 – Egypt/Xinhua: Peak accuracy of all models by number days in forecasting window

5.2.2 EXPANDING THE EVENT CATEGORIES

As discussed in the previous chapter, there are 20 basic classes of events under the CAMEO taxonomy, organized into four major Quad Classes: Material Conflict, Material Cooperation, Verbal Conflict, and Verbal Cooperation. The table and graph below show the breakdown of all Egyptian events identified by TABARI in Xinhua coverage during the six-year test and train periods into these four classes and the

total number of days that contain events from each of the categories. The graph displays the relative accuracy of the optimal model in forecasting each category of event. In both cases this includes the optimal model using a forecasting window of five days. It is clear that the majority of accuracy in forecasting general events comes from the high accuracy of Verbal Cooperation events and that the accuracy of each category appears to be related to the total number of training examples. Indeed, there is an r=0.99 Pearson correlation between the number of days exhibiting events in each category and the resulting maximal accuracy of the classifier in forecasting that event. Categories with more event-days might favor models with higher false positive rates. However, given that the accuracy metric used here is invariant to the raw number of low-event days, considering only the proportion of those correctly forecasted, this is largely accounted for. Instead, this likely reflects the fact that a higher number of training examples are able to more accurately capture the significant variation in the types of discourse that precede those classes of events, as classifiers are strongly sensitive to the breadth of the training text in terms of how closely it matches the test period text the model is eventually applied to. In addition, as explored later in this chapter, it is also possible that Verbal Cooperation events may be preceded by a more limited and regularized set of diplomatic terminology.

Indeed, a manual examination of a cross-section of Xinhua's coverage preceding Verbal Cooperation events illustrates that the majority appear to come from prescheduled scripted diplomatic meetings, such as a series of visits by President Mubarak to Italy ("Mubarak", 2004) or a summit in Sudan to broker a peace agreement between North and South Sudan ("Sudan", 2010). In each of these cases, there is often narrative for a week or more preceding the meetings that discusses the range of issues to be focused on at the meeting and potential outcomes. In this way, Verbal Cooperative events are rarely spontaneous or isolated: they are most often part of a broader sequence of diplomatic narrative justifying and outlining what each side hopes to accomplish at the meeting. Such meetings are often

preceded by physical diplomatic exchanges that would also be captured in the event stream and might offer event-based forecasting insights. However, the majority of the discourse preceding such events is speculative around what should be accomplished and thus only recordable through the narrative stream.

Of particular interest is that after Verbal Cooperation, the most accurate event classes are Material Conflict, Verbal Conflict, and Material Cooperation. That the two classes of conflict would be the next-most accurate forecasts likely reflects the long narrative lead-up to conflict that often occurs. Chadefaux (2012) found a strong upwards trend in conflict-related language 3-5 years before major conflict (Chadefaux, 2012), while Western Open Source Intelligence (OSINT) has long noted the "war of words" that often precedes major conflict (Lasswell, 1927; Roop, 1969).

Table 9 - Egypt/Xinhua: Peak accuracy at forecasting Quad Classes and their relative frequencies

Event Quad Class	Pos + Neg	Total Events	Days With Events	%Days With Events
allcount	131.72	24681	2045	82.96
quadmatconf	116.65	2472	746	30.26
quadmatcoop	114.17	1500	559	22.68
quadverbconf	115.27	1672	585	23.73
quadverbcoop	132.27	19037	1928	78.22

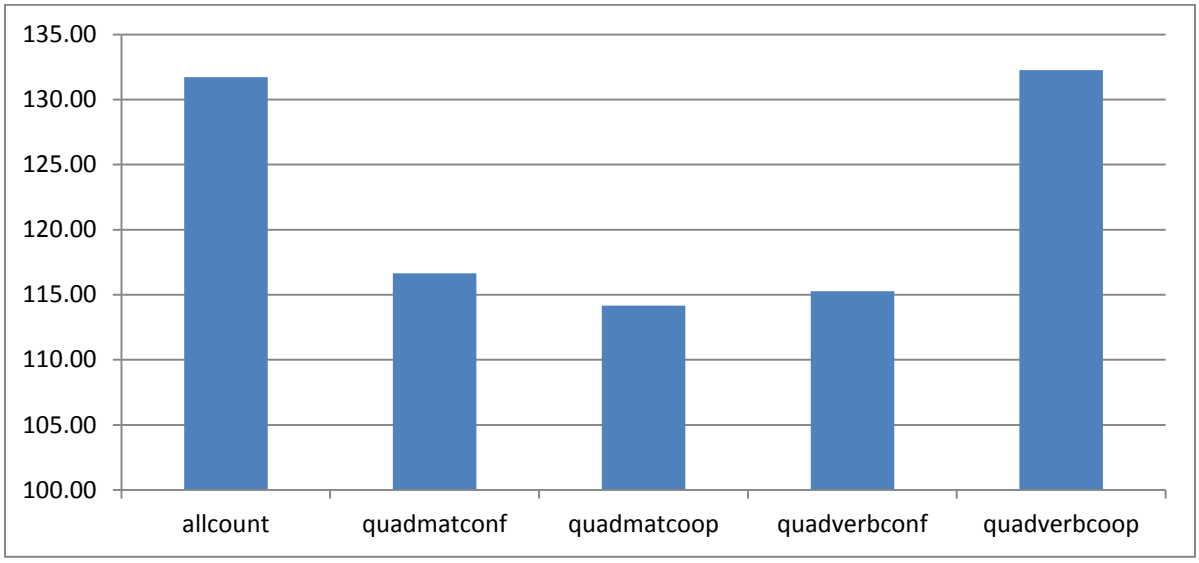

Figure 11 – Egypt/Xinhua: Peak accuracy at forecasting Quad Classes and their relative frequencies

5.2.3 USING MULTIPLE TEXT DAYS

Each of the examples above uses a single day of input text to forecast the number of events expected in subsequent days. Just as increasing the forecasting window boosted accuracy, it makes sense that increasing the number of days of input text might similarly lead to better accuracy. In particular, the more days a topic is discussed in the media, the more important it is likely to be, and expanding the input window to multiple days allows the model to take this information into consideration. The graph below shows the result of using META_GEO and increasing the number of input days from one day of text to using the preceding two, three, and four days of text to forecast the following day, two days, three days, four days, and five days of events. The most accurate model, achieving 43.59% better than random chance, uses the preceding two days of META_GEO text to forecast the number of events over the following four days. Intuitively it might seem that accuracy should increase with each additional day of input text, as it does for increases in the number of forecasted days. Instead there is substantial increase in accuracy when moving from a single day of text to two days of text, but accuracy then drops

when moving to three days, and four days of text actually results in lower accuracy than just the first day. This suggests that the strongest indicators occur in the 48 hours before an event, while increasing the window beyond this increases the noise level to a point where the strongest lexical features are drowned out by unrelated terms. Intriguingly, the work of Leskovec, Backstrom, and Kleinberg (2009) found that story lines tend to experience their most significant growth/decay cycle over precisely this 48 hour period.

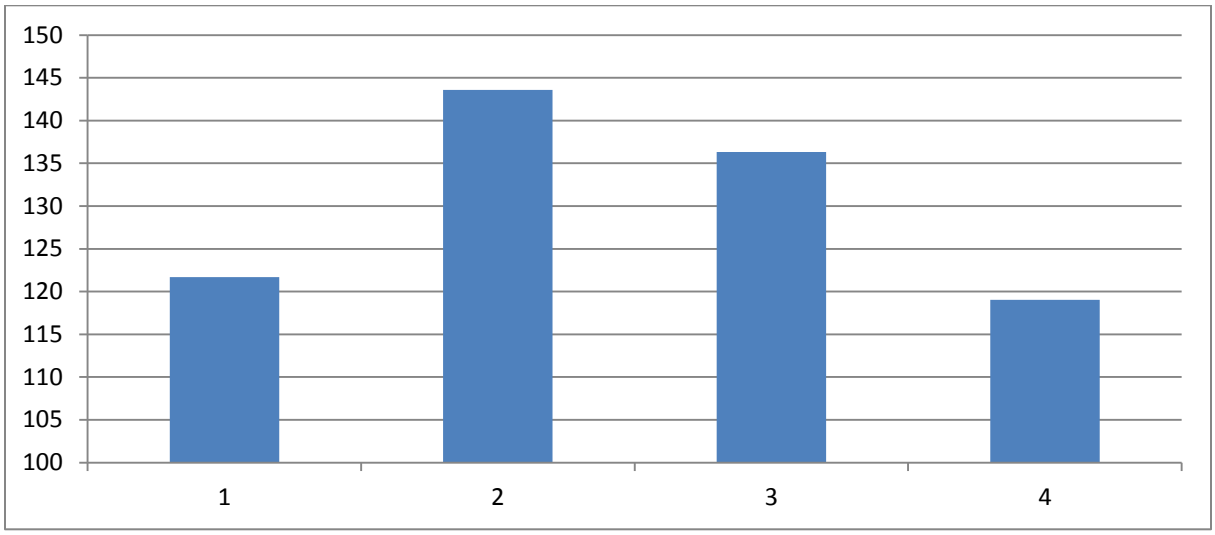

Figure 12 - Egypt/Xinhua: Peak forecasting accuracy by number of days of input text

5.2.4 NARROWING TO PROTESTS

There is an even split in the forecasting literature between forecasting broad classes of events like Shrodt (2000) and forecasting specific types of events like riots or disease outbreak (Radinsky & Horvitz, 2013). To test whether narrowing to just forecasting a specific type of event might be more accurate than attempting to forecast all types of events, the table below shows the results of examining only those events in "Class 1400 - Protests." During the combined training and testing period there were just 161 days with one or more Protest events, containing a total of 380 protest events, of which 103 of

those days (64%) occurred during the test period, totaling 286 protests. Protest events are highly bursty, with the majority of protests occurring in surges on isolated single days. In addition, the majority of protest events (75%) occur from December 2010 to March 2011, reflecting the unrest of the Egyptian revolution and represent a sharp departure from the day-to-day norm of the training period. Using two days of input text and forecasting the following day's events using a threshold of one event, accuracy is fairly low, reflecting the sharp departure of these protests from the previous periods. The extreme false positive rate indicates that the models simply could not find sufficient distinguishing language.

Most intriguing is that the only text type to achieve measurable accuracy is FTXT_TONEPOS, suggesting that surges in positive language are most predictive of protests during this period. Indeed, while the protests involved violent clashes and destruction of property ("18 Days", 2011; "Protests", 2011), there was significant optimism portrayed in coverage of the protests, part of the broader discourse around the growing Arab Spring, with leaders "look[ing] ahead towards the future" and dreaming of a "national consensus that would… guarantee the interests of [citizens]" ("AL", 2011). Egypt even asserted it would "follow the developments in Tunisia closely and will respect people's choice" and "what they have achieved since their independence" ("Respect", 2011). Tunisians, in return, rallied in mass gatherings in support of their Egyptian counterparts ("Rally", 2011). Egypt even played a critical role as an evacuation staging point during the subsequent unrest in Libya, with many countries evacuating their citizens through embassies and ports in the neighboring country ("Embassy", 2011).

Table 10 - Egypt/Xinhua: Accuracy at forecasting Protest events

Text Type	True Pos	True Neg	Accuracy	Pos + Neg
FTXT_TONEPOS	96.10	4.95	8.08	101.05
META_GEO	100.00	1.02	4.42	101.02
FTXT_TABARIALLACTORS	100.00	0.18	3.62	100.18
FTXT_TONE	100.00	0.14	3.57	100.14
FTXT_RELIGIOUS	100.00	0.13	3.97	100.13
FTXT_POSNONOUNS	100.00	0.09	3.53	100.09
FTXT_STOPWORDS	100.00	0.05	3.48	100.05
FTXT_STEMMED	100.00	0.05	3.48	100.05
FTXT_CLEAN	100.00	0.05	3.48	100.05
FTXT_POSNOPROPERNOUNS	100.00	0.05	3.48	100.05
FTXT_TONENEG	100.00	0.05	3.48	100.05
FTXT_TABARIVERBS	0.00	100.00	96.56	100.00
FTXT_POSVERBS	0.00	100.00	96.56	100.00
META_ORG	0.00	100.00	96.56	100.00
FTXT_POSADJS	0.00	100.00	96.56	100.00
FTXT_ETHNICRELIGIOUS	0.00	100.00	96.55	100.00
FTXT_ETHNIC	0.00	100.00	96.54	100.00
FTXT_STEMMEDANDSTOPWORD	0.00	100.00	96.56	100.00
META_NAME	0.00	100.00	96.56	100.00

5.2.5 PEERING INSIDE THE MODELS

To better understand the specific language being focused on by the models above, Table 11 reports the conditional probabilities learned by the FTXT_TONEPOS model discussed above for Egyptian Protest events in Xinhua. This represents the word weightings learned by the model from the training data that are most or least predictive of future Protest events. In all, just 29 words appeared with sufficient regularity to be incorporated into the model, while just a single word, "peace," was found to be predictive of protests. At first glance this would appear counterintuitive, however upon review of Xinhua's coverage of Egyptian protest events, Xinhua appears to focus on calls by authorities for protesters to return home or to keep the peace during periods of instability. For example, after a

massive riot of 5,000 people on October 21, 2005 at St. George's Coptic Church in Alexandria, in which more than 100 people were injured, the "Grand Imam of Al Azhar and the Pope of Alexandria of Egypt called on Saturday for both Muslims and Coptic Christians in Egypt to appeal to reason and keep peace and security" ("Leading", 2005). In fact, this is a common theme throughout Xinhua's coverage of protest: in the days preceding actual protest events, there are calls from authorities for "peace," which are eventually ignored. The presence of terms like "agreement," "ceremony," and "diplomatic" appear to have been learned by the model to appear near mentions of "peace" with respect to forthcoming diplomatic meetings, such as peace summits, biasing away from Protest events.

Table 11 - Egypt/Xinhua: FTXT_TONEPOS model term conditional probabilities for protest events

Term	Low Mean	Low Var	High Mean	High Var
agreement	5.51	12.60	0	0
peace	4.38	4.22	4.99	1.01
support	3.30	8.31	0	0
party	3.02	10.05	0	0
agreed	2.24	10.81	0	0
free	2.10	7.86	0	0
health	2.00	8.36	0	0
ambassador	1.88	7.42	0	0
fair	1.69	6.46	0	0
agreements	1.59	7.11	0	0
ceremony	1.41	14.19	0	0
prince	1.33	6.51	0	0
aid	1.12	5.63	0	0
special	1.08	6.51	0	0
liberation	0.97	6.68	0	0
unity	0.90	5.21	0	0
safe	0.87	5.84	0	0
natural	0.79	5.48	0	0
solidarity	0.75	5.97	0	0
good	0.73	5.06	0	0
brotherhood	0.68	5.46	0	0
diplomatic	0.59	4.54	0	0
crown	0.54	4.13	0	0

Table 11 (cont.)

Term	Low Mean	Low Var	High Mean	High Var
neighbors	0.54	4.28	0	0
relief	0.52	5.18	0	0
strategic	0.51	5.17	0	0
unified	0.48	4.64	0	0
confidence	0.46	4.44	0	0
helping	0.39	3.70	0	0

The table below shows a sample of the raw conditional probability output of the FTXT_TONEPOS model for 11 days. These are the respective probabilities that the following day will belong to either the Low Event or High Event categories, respectively. It is immediately clear that the majority of the false positive forecasts are through a significant margin (median of 6.81E-89 difference between High and Low), suggesting that the model is significantly off. However on day 472 the difference fell to 0.18E-19, suggesting that some additional tuning of the model could potentially increase the accuracy slightly.

Table 12 - Egypt/Xinhua: Raw conditional probability output by day

Day #	Low Prob	High Prob	Predicted	Actual
468	2.75463E-94	1.00E+00	High	Low
469	2.19E-94	1.00E+00	High	Low
470	3.45E-88	1.00E+00	High	Low
471	2.75E-97	1.00E+00	High	Low
472	1.18E-19	1.00E+00	High	Low
473	7.48E-113	1.00E+00	High	Low
474	1.00E+00	4.34E-12	Low	Low
475	2.82E-94	1.00E+00	High	Low
476	6.81E-89	1.00E+00	High	Low
477	2.44E-69	1.00E+00	High	Low
478	6.70E-73	1.00E+00	High	Low

Returning to Table 8 from earlier in this chapter, the META_GEO surrogate achieved just over 11% greater accuracy than random chance when forecasting the incidence of any type of event the following

day. Table 13 shows the top ten terms from the model's conditional probability table that had the highest bias towards High Event days. The first and third terms refer to the "Gaza Strip," while the tenth term also relates to the Israeli-Palestinian conflict. Egypt played a significant diplomatic role in the Israeli-Palestinian conflict during this period, especially through a series of discussions with the Italian government, which "underlined the need for the European Union to support Egypt's efforts to help the Palestinians prepare for an expected Israeli withdrawal from Gaza" ("Visit", 2004). Egypt had extensive relations with Italy at the time, signing a "Cultural and Archaeological Cooperation protocol for exchanging expertise and archaeological missions" ("Leading", 2004) and instantiating a formal bilateral security cooperation agreement ("Agreement", 2004). Italy was even viewed as a utopia for many, with "migration to Italy … becom[ing] a dream for many Egyptian youth, particular those residing in the Nile Delta governorates" ("Deports", 2005). Indeed, according to a 2010 article "Italy-Egypt ties are living a flourishing moment… the agreements signed on Wednesday further enhanced the Italy-Egypt strategic partnership frame launched in 2008" leading to a "special relationship" ("Strategic", 2010). The tie with Mozambique appears to come from Egypt often being discussed alongside other African nations due to its being grouped by major Non-Governmental Organizations as part of Africa rather than the Middle East, as in a 2003 report on African economic progress that studied Egypt alongside 6 other African nations, discussing the "deteriorating political and economic situation" in several of them ("Performers", 2003). Egypt's connection with Bangladesh appears to come through its membership in the D-8 group of nations (Egypt, Bangladesh, Indonesia, Iran, Malaysia, Nigeria, Pakistan, and Turkey) ("Redoubling", 2001).

That META_GEO is predictive of future events is therefore unsurprising, as it suggests that specific locations play predefined roles in Egyptian society at different times. Any discussion of Italy in news coverage of Egypt during this period is likely to revolve around a new partnership between the two

nations, while discussions of the Gaza Strip or Israel are likely to involve Egyptian attempts at brokering peace. This largely reflects Egypt's role as a regional diplomatic courier, with such diplomatic partnerships (or their precursors such as the fallout of a recent conflict) likely formally scheduled and discussed over the days and weeks leading up to an event. The connections to Mozambique and Bangladesh are similarly due to the way in which Egypt is viewed by the world (as part of Africa) and the way it views itself in the world through the regional partnerships it joins (the D-8 group).

Yet, beyond its practical applicability to forecasting, the situating nature of these spatial ties also reflects a strong geographical dimension to cultural framing: that of "cultural geography," which places "individual actors and isolated communities… [into the] shifting web of interdependencies which often stretches across the globe" (Jackson, 1989). While cultural geography traditionally refers to the way in which culture varies across geography, here it is reflective of how specific geographic locations can occupy highly specific and predictive roles in a society. Of course, the term-based Naïve Bayesian models used here can only offer lists of locations and their associated probabilities with respect to future events. The contextualization of those locations into a descriptive narrative tying them into a broader cultural framing must therefore be deferred to the human user. Yet, even so, the ability of this approach to guide a human user to unexpected geographic ties and to provide quantitative affirmation of expected ties, offers significant potential for visualizing, exploring, and understanding these patterns in more detail.

Table 13 - Egypt/Xinhua: Model conditional probabilities for META_GEO: locations most predictive of High Event days

Term	Low Mean	Low Var	High Mean	High Var	High Bias
strip	5.51	11.71	9.50	19.67	3.99
syria	6.19	10.72	8.33	12.60	2.14
gaza	6.98	13.09	8.95	18.70	1.97
italy	2.13	6.91	4.05	11.02	1.92
bangladesh	0.18	2.50	1.83	8.36	1.65
mozambique	0.32	3.75	1.82	8.31	1.49
damascus	0.23	3.21	1.39	8.35	1.16
germany	2.67	9.31	3.53	10.53	0.86
columbia	0.31	3.67	1.13	8.11	0.81
jerusalem	1.79	7.40	2.60	9.55	0.81

To explore this concept further, the table below shows the same model, but displays the top ten terms that most strongly bias towards a Low Event day, rather than a High Event day. Surprisingly, the top three terms once again relate to the West Bank and Palestinians. Delving further, it appears the driving force between this dualism of linguistic indicators is that Xinhua describes the contested region as the Gaza Strip when there is high potential for an event the following day, but discusses it as the West Bank or to the people living there as Palestinians when there is a much lower probability of an event. Indeed, a closer analysis of articles during the training period illustrates that articles referring to Gaza often document specific ongoing activities, such as Egyptian border guards killing smugglers at the Egyptian-Gaza border ("Smugglers", 2005), while articles referring to the Palestinian people refer more often to diffuse summaries that do not constitute discrete codeable events, such as "continuing [Israeli] military activities in the Palestinian territories" ("Truce", 2005). This illustrates a fundamental concept of media framing in journalism: different terms can have specific connotations or be used in different contexts to reflect different approaches to an issue (Tankard, 2001). It also suggests that models are likely to be the most accurate when they access the original wording, rather than applying synonym translation or other normalization approaches (Rodriguez, Hidalgo & Agudo, 2000).

Table 14 - Egypt/Xinhua: Model conditional probabilities for META_GEO: locations most predictive of Low Event days

Term	Low Mean	Low Var	High Mean	High Var	High Bias
bank	10.87	16.60	5.22	9.76	-5.65
west	10.83	16.52	5.23	9.71	-5.61
palestinian	5.41	8.73	0.28	1.83	-5.13
saudi	9.05	12.40	4.71	8.77	-4.34
jordan	9.19	13.27	4.93	9.45	-4.26
arabia	8.51	12.01	4.61	9.34	-3.90
general	5.36	8.48	1.84	5.35	-3.52
morocco	5.16	14.26	1.65	5.21	-3.51
ramallah	3.76	12.29	0.55	4.62	-3.21
yemen	4.26	10.64	1.23	5.31	-3.04

5.2.6 LEARNING FROM MORE RECENT KNOWLEDGE

Each of the models above learns from six years of news about Egypt (1999 through 2005) and applies that learned knowledge to forecast events for the following six years (2006 through the end of 2011). Yet, the test period contains the Arab Spring and a national revolution within Egypt that brought down its government, while the training period contains no incidents that approach this level of societal unrest. Indeed, a paper published by the administrator in charge of ICEWS a year before the Arab Spring noted the reliance of many forecasting models on studies of decades-old conflicts that might not reflect modern realities (O'Brien, 2010).

To test this theory, the same modeling process is repeated, but here the training period is limited to November 1, 2008 (since September 2008 Xinhua content is unavailable in LexisNexis) through October 31, 2010, with the test period running from November 1, 2010 through August 31, 2012. This results in an approximate balance of two years of training data to forecast the following two years, vastly reducing the time horizon over which the models must function from 12 years down to 4 years. As before, an

array of models are tested, exploring High Event thresholds of the mean, median, first, and third quartiles of the number of events per day in the training period. In each case, the preceding two days of text are used to forecast whether the following day will be a High or Low Event day.

The table below summarizes the results of this analysis, demonstrating the most accurate forecasts yet, approaching even the accuracy of the relaxed five-day forecasting window, but when using the more stringent single-day criterion. The most accurate text type, FTXT_POSADJS, achieves accuracy of 25% better than random chance, with 77% accuracy at recognizing Low Event days and nearly 50% accuracy at recognizing High Event days, placing it within the realm of possible actionable use (O'Brien, 2010). In practice, such a system might update its internal model each day, retraining on a rolling window of the past two years, eliminating older knowledge and constantly incorporating new knowledge into the model. The model could also be weighted to place more emphasis on more recent articles, biasing its knowledgebase towards more recent events.

Of additional note, the third most accurate surrogate is FTXT_CLEAN, the original raw unfiltered text. This is significant in that it means that the most basic model, without the benefit of the theoretic underpinnings that drive the filtered versions of the text, can achieve substantial accuracy at forecasting future conflict. Given there are likely substantial accuracy improvements that could be achieved through more advanced modeling techniques, it is likely that with further work, this accuracy could be improved even further, and opens the possibility of using basic classifiers to operationally predict conflict in novel situations.

Table 15 - Egypt/Xinhua best models (training 10/2008 – 10/2010; testing 11/2010 – 8/2012)

Text Type	High Thres	True Pos	True Neg	Accuracy	Pos + Neg
FTXT_POSADJS	9	47.71	77.78	65.97	125.49
FTXT_STEMMED	9	56.87	67.41	63.27	124.28
FTXT_CLEAN	9	53.82	68.64	62.82	122.46
FTXT_POSNONOUNS	9	50.00	72.10	63.42	122.10
FTXT_POSNOPROPERNOUNS	9	51.91	69.88	62.82	121.78
FTXT_STOPWORDS	9	50.00	71.36	62.97	121.36
FTXT_POSVERBS	14	48.10	73.08	67.17	121.19
FTXT_TONENEG	9	58.78	61.98	60.72	120.75
META_GEO	9	59.16	60.49	59.97	119.65
FTXT_STEMMEDANDSTOPWORD	9	50.00	69.63	61.92	119.63
FTXT_TABARIVERBS	2	35.14	84.02	47.53	119.16
FTXT_TABARIALLACTORS	2	37.35	81.07	48.43	118.41
FTXT_ETHNIC	10	40.42	77.96	64.35	118.38
META_NAME	14	27.22	89.00	74.36	116.21
FTXT_TONE	9	51.91	63.95	59.22	115.86
FTXT_TONEPOS	6	42.26	72.81	57.42	115.07
META_ORG	6	32.14	81.82	56.76	113.96
FTXT_ETHNICRELIGIOUS	14	29.11	81.69	69.22	110.81
FTXT_RELIGIOUS	3	7.27	96.45	34.48	103.72

The table below examines FTXT_CLEAN in more detail, listing the top 10 terms most closely associated with subsequent High Event days. It is of interest that four of the ten are location-related. Sharm and Sheikh refer to Sharm el-Sheikh, which is a popular destination for regional peace conferences. Thus, the system has learned that increases in discussion of this city likely indicate an impending peace conference, which ties in with the most strongly weighted term, "summit." The focus on ceasefire, offensive, and strikes reflect heavy clashes between Israeli and Palestinian forces during the training period Xinhua coverage invariably noted that these clashes occurred after "Egypt failed to renew a ceasefire between Israel and the Palestinian factions" ("Militants", 2009), framing the resulting conflict as something Egypt could have prevented. Indeed, the term "ceasefire" appears driven by Hamas' alleged propensity to fire rockets from Egypt into Israel and the resulting protests, diplomatic

exchanges, and discussions therein ("Eilat", 2010). The dialogues and meetings during this period regarding the Israeli-Palestinian conflict brought Egypt together with Kuwait, Qatar, and other regional nations ("Patch", 2009), including Kuwait's backing for Palestinian statehood at meetings to be held in Egypt later in 2009 ("Reaffirms", 2009). This reinforces Egypt's significant interplay with regional Verbal Cooperation and Material Conflict events during this period and the resulting high ability to forecast such events.

Table 16 - Egypt/Xinhua: FTXT_CLEAN conditional probabilities (training 10/2008 – 10/2010; testing 11/2010 – 8/2012)

Term	Low Mean	Low Var	High Mean	High Var	High Bias
summit	7.50	35.17	17.64	51.21	10.13
ceasefire	6.16	25.39	14.13	41.50	7.97
aligned	1.94	26.98	9.29	51.24	7.35
offensive	1.27	7.49	6.57	19.22	5.30
strikes	0.57	5.27	5.62	30.09	5.05
sharm	5.62	29.06	10.51	41.48	4.90
sheikh	5.54	26.35	10.33	38.63	4.79
kuwait	1.08	8.72	5.74	27.28	4.66
emergency	1.35	6.93	6.00	27.19	4.66
doha	2.13	24.22	6.66	33.42	4.53

Table 15 epresents a significant increase in accuracy over the original 12-year model results of Table 5. Yet, rather than being a result of more recent training data yielding better results, this could merely be a result of less input text, which reduced the amount of noise words in the model. Indeed, reducing the input text from using full text to lead paragraphs significantly increased accuracy in a similar way. Thus, the table below repeats this experiment, using an earlier training period of November 1, 2002 through October 31, 2004, while keeping the test period the same. The resulting models are even more accurate than in Table 15 and there is yet another reordering of the text types.

Table 17 – Egypt/Xinhua best models (training 11/2002 – 10/2004; testing 11/2010 – 8/2012)

Text Type	High Thres	True Pos	True Neg	Accuracy	Pos + Neg
FTXT_ETHNIC	14	58.23	71.23	68.13	129.46
FTXT_ETHNICRELIGIOUS	14	61.39	67.91	66.37	129.31
FTXT_POSVERBS	6	48.21	74.92	61.47	123.14
FTXT_CLEAN	11	74.18	48.46	56.67	122.64
FTXT_STEMMEDANDSTOPWORD	14	82.28	38.90	49.18	121.18
FTXT_POSNONOUNS	11	77.00	43.17	53.97	120.17
FTXT_STOPWORDS	11	80.28	39.87	52.77	120.15
FTXT_TONEPOS	14	53.80	65.62	62.82	119.42
META_GEO	11	70.42	48.68	55.62	119.10
FTXT_TONE	6	42.56	75.53	58.92	118.09
META_NAME	14	31.65	86.25	73.31	117.89
FTXT_STEMMED	11	71.36	46.26	54.27	117.62
FTXT_POSADJS	6	69.05	48.04	58.62	117.08
META_ORG	11	36.15	78.81	65.17	114.96
FTXT_TONENEG	6	30.65	83.38	56.82	114.04
FTXT_RELIGIOUS	20	34.12	79.74	72.74	113.86
FTXT_POSNOPROPERNOUNS	11	72.30	40.53	50.67	112.83

Examining the conditional probability table of FTXT_CLEAN for this model, “summit” and “crisis” again are top terms, but the list also reflects a focus on Egypt’s public support for Sudan and Darfur (“Darfur”, 2004), while the discussion of the Red Sea and foreigners was driven by a set of bomb attacks against resorts in the area which killed many foreign tourists (“Condemns”, 2004). Turkey’s prominence stemmed from close relations between the two countries during this period, including an agreement on Palestine (“Consensus”, 2004).

Table 18 - Egypt/Xinhua term conditional probabilities (training 11/2002 – 10/2004; testing 11/2010 – 8/2012)

Term	Low Mean	Low Var	High Mean	High Var	High Bias
summit	4.63	16.19	7.54	19.70	2.91
arab	5.58	11.06	8.09	14.56	2.50
prime	1.63	6.05	4.04	8.54	2.41
crisis	1.91	7.61	4.23	13.34	2.32
sea	1.23	8.58	3.48	17.31	2.25

Table 18 (cont.)

Term	Low Mean	Low Var	High Mean	High Var	High Bias
foreign	4.03	6.80	6.21	8.60	2.18
meeting	3.61	8.12	5.71	10.74	2.10
red	1.17	7.96	3.21	16.07	2.04
turkey	0.81	6.44	2.84	10.36	2.03
darfur	0.48	4.34	2.27	23.28	1.79

Repeating this again for the same test period, but using the period 11/2/2005 to 11/1/2007 as the training period, the order of the text inputs changes again, but this time accuracies are more on par with the original training period from Table 15.

Table 19 - Egypt/Xinhua best models (training 11/2005 – 10/2007; testing 11/2010 – 8/2012)

Text Type	High Thres	True Pos	True Neg	Accuracy	Pos + Neg
FTXT_POSNOPROPERNOUNS	10	71.67	51.05	58.47	122.72
FTXT_POSNONOUNS	8	74.56	47.89	59.37	122.46
FTXT_POSVERBS	10	65.42	55.27	58.92	120.69
FTXT_TABARIALLACTORS	4	73.15	47.51	63.12	120.66
FTXT_CLEAN	8	64.46	56.05	59.67	120.51
FTXT_TONENEG	10	69.58	50.82	57.57	120.40
FTXT_ETHNIC	16	42.31	77.44	70.54	119.75
FTXT_STEMMEDANDSTOPWORD	4	42.12	77.39	55.92	119.51
FTXT_POSADJS	4	52.71	66.67	58.17	119.38
FTXT_STEMMED	4	68.72	50.57	61.62	119.29
FTXT_STOPWORDS	8	64.11	54.74	58.77	118.85
META_GEO	10	79.17	39.34	53.67	118.51
FTXT_ETHNICRELIGIOUS	16	38.46	79.48	71.47	117.94
FTXT_TABARIVERBS	10	76.25	41.69	54.12	117.94
META_NAME	10	28.33	89.46	67.47	117.79
FTXT_TONE	8	60.98	55.26	57.72	116.24
FTXT_TONEPOS	8	42.51	70.79	58.62	113.30
META_ORG	10	18.33	91.78	65.32	110.12
FTXT_RELIGIOUS	16	83.33	18.86	32.13	102.20

The table below once again shows the top ten terms from FTXT_CLEAN that bias towards High Event days. In this case, the prevalence of "plane" appears driven by the air evacuation of Egyptian citizens from Lebanon in July 2006 ("Nationals", 2006). At the time there were at least two prominent Egyptian political leaders with the name Ahmed, including People's Assembly Speaker Ahmed Sorour ("Speaker", 2007) and Egyptian Foreign Minister Ahmed Abul Gheit ("Consultations", 2007), both of whom were widely involved in Verbal Cooperative agreements. France's prevalence appears driven by strong diplomatic relations between it and Egypt over the Hezbollah-Israeli conflict ("FM", 2006).

Table 20 - Egypt/Xinhua term conditional probabilities (training 11/2005 – 10/2007; testing 11/2010 – 8/2012)

Term	Low Mean	Low Var	High Mean	High Var	High Bias
plane	0.00	0.00	2.67	25.55	2.67
chief	1.02	4.02	3.63	8.50	2.60
saudi	2.69	9.42	5.19	13.25	2.51
talks	3.23	6.39	5.63	7.98	2.40
ahmed	2.01	5.60	4.37	8.48	2.36
french	0.93	5.99	3.09	14.72	2.16
prime	2.57	7.04	4.71	8.17	2.14
african	0.72	6.11	2.63	9.44	1.91
arab	4.99	7.54	6.86	11.26	1.87
foreign	3.63	5.32	5.42	6.15	1.79

Perhaps most intriguing about this series of experiments is that as the training period is moved backwards from the test period, the accuracy of the models does not decrease solely with age, as might be expected if information had a natural lifespan associated with it that decayed with time irrespective of all other factors. Empirical work on media cycles such as Leskovec, Backstrom, and Kleinberg (2009) have long demonstrated that individual storylines have relatively short lifespans, usually measured in days to hours and that newer information continuously displaces older information. This is largely due to the physical constraints that historically limited the volume of information that a news outlet could

disseminate each day, known as the "newshole" (Gans, 1979). A print newspaper has only so many words it can print each day, a television or radio news station is limited to the 24 hours in each day, and even web publications are limited by the volume of content their reporters can write.

At the same time, the dual concepts of "recency and novelty" form two of the cornerstones of "newsworthiness" in which the newer a piece of information is, the more "worthy" it is of reporting (Galtung & Ruge, 1965; Golding & Elliott, 1979; Bell, 1991). Thus, if the newest information holds the greatest "value" and the fixed size of the newshole means that covering older events would displace this newer information, the news media becomes a temporally-constrained window over societal behavior that covers situations as quickly as possible after they occur and phases them out of discussion almost as quickly (Doran, Pendley & Antunes, 1973; Moeller, 1999). Furthermore, in news-driven fields such as economics, information is most valuable when it is in limited circulation and thus the competitive advantage or "value" of a given story can be diminished even while it is still receiving media attention (Stierholz , 2008; Halzack, 2012).

If newer information is the most "newsworthy" and displaces older information as it arrives, with older information losing economic value with the length of time it has been in circulation, this would suggest that information should experience an "aging" process in which it becomes less valuable over time. In fact, however, simply because a story has ceased to be covered by the news media does not mean its societal impact has been diminished. In fact, the "impact" of an event and its "newsworthiness" are largely decoupled processes (Doran, Pendley & Antunes, 1973; Moeller, 1999). This is reflected in the peak model accuracies seen above, in which the oldest training period, ending six years before the training period, yields an accuracy of 29% better than random chance, followed by 22% greater accuracy three years prior and 25% better immediately prior. While this still represents significant change in

accuracy over the three periods, accuracy does not decrease directly with age: in fact the oldest training period yields the highest accuracy. This suggests an alternative explanation, that rather than decaying directly with time, the predictive power of information changes in concert with how closely the underlying events and discourse of the training period aligns with the testing period. This is in following with the general literature on text classification systems that models yield the highest accuracy when the training and testing periods are as closely aligned as possible (Sebastiani, 2002).

In the context of the predictive value of information, this suggests that beneath the collection of specific time-dependent storylines covered by the news media each day exist underlying time-independent (or at least long horizon) master narratives or "metanarratives" into which those shorter narratives are situated (Lyotard, 1984). For example, while a mass shooting in the United States may fade from the news agenda within a few days, it is situated in a time-indefinite American metanarrative of gun ownership (Lott, 2010) that sustains long-term thematic continuity across the individual discussions of specific shootings. Indeed, the continuity of metanarratives and their interplay with time-dependent stories are a key component of the cultural framing that both influences and is influenced by the information sphere (Gupta, 1992).

Turning to the geographic locations most closely associated with Egyptian events in each training period, in 2002-2004 it was Turkey, the Red Sea, and Darfur, in 2005-2007 it was France, and in 2008-2010 it was Sharm el-Sheikh, Kuwait, and Doha. That there would be such disunity in locative contextualization across the three time periods, yet each achieved relatively similar accuracy at forecasting the same future two year period, suggests each of those locations instantiates a story contained within a larger metanarrative of Egypt as a regional superpower (Barnett, 1993). Throughout all three training periods Egypt played a central role in mediating regional conflicts, especially clashes between Israel and

Palestine. In this way, while the specific countries Egypt interacts with at any moment may change from month to month, its role as a peace broker and regional diplomat remains constant over this period. Thus, even as each country's prominence in media coverage of Egypt declines, its role with respect to Egypt's activities remains stable enough to be predictive of future events. This may largely be due to the relative stability imposed by Egypt's long standing dictatorship and the relatively short analytical period of just the past decade examined here. Yet, even if these factors were significant, it still demonstrates that at least in a relatively unchanging country, the predictive value of information appears less influenced by its age and more by the degree to which it aligns with the test period.

5.3 EXPANDING TO OTHER SOURCES AND COUNTRIES

Studying Egyptian events through the eyes of Xinhua suggests a number of key traits of information and discourse as they relate to forecasting. Yet, to ensure such findings are generalizable, it is critical to expand the analysis to consider additional news sources and additional countries, especially from across different geographic regions. Four additional countries (Indonesia, South Africa, Brazil, and Germany) from two additional sources (Agence France Presse and Associated Press) are added to the Egypt/Xinhua analyses.

Indonesia was selected for its long history of ethnic and religious conflict (Bertrand, 2004), ranking 24th on Fearson's (2003) Ethnic Fractionalization Index. This provides an ideal test of whether the ethnic and religious group-based text filters properly reflect such group-based discourse in countries with known group-based conflict. It is also the 19th-most mentioned country in Xinhua, 17th in Agence France Presse, and 40th in Associated Press. This suggests there is sufficient coverage volume to offer a reasonable

number of input documents to the models, compared to a territory like the Midway Islands, with just 21 articles mentioning it in Xinhua over the entire 1979 to 2012 period. As an Asiatic country, Indonesia allows the exploration of regional differences in coverage and its close proximity to China tests whether Xinhua covers Asiatic or neighboring countries differently.

Representing Africa, South Africa was selected as the African nation with the most coverage across the three sources, the 16th-most covered country overall in Xinhua, 29th in Agence France Presse, and 23rd in the Associated Press. As a native English-speaking country and the location of the Associated Press' bureau for all of southern Africa ("Torchia", 2013), it also offers the ability to explore whether this increases Associated Press attention of the country due to greater cultural linkages (Kareil and Rosenvall, 1984; Wu, 2006). This could also indirectly affect Xinhua and Agence France Presse coverage of South Africa, even though English is not their primarily publication language, in that greater global coverage could indirectly feed back into the attention paid to the nation, increasing its news value.

Brazil was selected to represent Latin America as the world's fifth largest country in terms of geographic area and population. It is ranked 40th in Xinhua, 58th in Agence France Presse, and 45th in Associated Press coverage. Finally, in Europe, France is the most-discussed continental European nation in all three sources, but given that Agence France Presse is based there, this could potentially bias its findings, even in spite of the filters used to identify only international stories from the newswire. Thus, Germany, the next-most-common European nation was selected, ranked 12th in Xinhua, 8th in Agence France Presse, and 6th in the Associated Press. Germany's prominent international role as a financial center could also shape its coverage in the three sources (Wu, 2006).

5.3.1 EXPLORING THE COUNTRIES

Table 21 shows the total number of articles mentioning each country in the training and testing periods, the total number of days containing one or more articles in each period, and the total number of events of any kind in each period. The training period is set to the same November 1, 2008 - October 31, 2010 period used earlier, with the same November 1, 2010 - August 31, 2012 testing period. Table 22 breaks out the number of events in each period into their respective Quad Class categories of Material Conflict, Material Cooperation, Verbal Conflict, and Verbal Cooperation. The last column for each Country/Source entry shows the Pearson correlation between the event distribution for the training and testing periods, with all falling into the range r=0.98-0.99. This illustrates the training and testing periods are closely aligned in terms of the relative distributions of event classes.

Table 21 – Total number of articles and events by train and test periods for all countries and sources (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Country	Source	Train Arts	Test Arts	Train Days	Test Days	Train Events	Test Events
germany	afp	28162	29501	730	670	12790	12039
egypt	afp	11304	22463	729	670	9357	15852
egypt	xinhua	19673	20945	725	668	7251	7381
indonesia	xinhua	13634	11113	719	661	3457	3118
germany	xinhua	12097	10061	718	659	2782	2359
southafrica	afp	6104	7946	720	668	1450	1713
egypt	apmain	4525	6992	730	670	2533	5063
brazil	afp	7527	6896	717	665	4464	3000
indonesia	afp	7589	6138	729	666	4941	3536
germany	apmain	6446	5857	730	670	3323	2317
southafrica	xinhua	5282	5446	684	635	929	927
brazil	xinhua	9789	4849	709	615	3617	1354
southafrica	apmain	2519	2650	730	670	425	460
brazil	apmain	2889	2513	729	670	1406	636
indonesia	apmain	2334	2409	730	670	913	824

able 22 – Breakdown of event types in training and testing periods by country and source (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Country	Source	Train-MatConf	Train-MatCoop	Train-VerbConf	Train-VerbCoop	Test-MatConf	Test-MatCoop	Test-VerbConf	Test-VerbCoop	Event Correlation
germany	afp	1612	1043	1485	8650	1394	870	1383	8392	0.9999
egypt	afp	1139	410	904	6904	2568	1233	2476	9575	0.9976
egypt	xinhua	513	339	478	5921	1195	735	743	4708	0.9954
indonesia	xinhua	430	284	250	2493	353	417	168	2180	0.9946
germany	xinhua	337	208	236	2001	218	211	189	1741	0.9984
southafrica	afp	219	119	168	944	295	106	153	1159	0.9977
egypt	apmain	314	379	270	1570	889	758	768	2648	0.9945
brazil	afp	376	287	347	3454	389	279	324	2008	0.9995
indonesia	afp	1184	355	588	2814	776	282	441	2037	0.9982
germany	apmain	557	493	400	1873	352	313	260	1392	0.9997
southafrica	xinhua	104	49	77	699	104	50	87	686	0.9999
brazil	xinhua	221	196	259	2941	89	96	75	1094	0.9993
southafrica	apmain	63	73	42	247	63	82	36	279	0.9993
brazil	apmain	144	234	167	861	118	106	78	334	0.9833
indonesia	apmain	244	192	75	402	204	127	56	437	0.9794

Table 23 ranks all text types by their peak and average accuracies across all countries and news sources, while Table 24 breaks the results down by news source. Agence France Presse exhibits such high accuracies across all text types that it ends up dominating the majority of the entries in Table 23. In fact, as Table 25 shows, it is the most accurate news source overall across all five countries, 8% more accurate overall than Xinhua, which, in turn is 9% more accurate than the Associated Press. Table 9 showed that in Xinhua coverage of Egypt, Verbal Cooperation events were vastly more predictable than other types of events. Verbal Cooperation events constitute 74% of all Xinhua events for the five countries, 66% of Agence France Presse events, and 56% of Associated press events, so the higher accuracy of Agence France Presse does not appear due to it having a higher density of more predictable events and thus appears to be an actual characteristic of the outlet itself.

Table 23 – Peak accuracy by text type across all countries and sources (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Text Type	Max Pos + Neg	Average Pos + Neg
FTXT_CLEAN	135.54	115.92
FTXT_POSNOPROPERNOUNS	134.06	115.84
FTXT_POSVERBS	131.57	115.67
FTXT_POSADJS	130.61	115.39
FTXT_POSNONOUNS	130.13	115.17
FTXT_STEMMED	133.83	115.07
FTXT_STEMMEDANDSTOPWORD	132.01	115.02
FTXT_STOPWORDS	133.98	114.56
FTXT_TABARIALLACTORS	132.71	113.93
FTXT_TABARIVERBS	133.51	113.86
META_GEO	128.89	112.36
FTXT_TONE	131.35	111.97
FTXT_ETHNIC	122.30	111.89
FTXT_TONENEG	131.96	111.71
FTXT_ETHNICRELIGIOUS	118.53	110.87
FTXT_TONEPOS	132.05	110.82
META_NAME	123.81	110.23
META_ORG	129.11	108.93
FTXT_RELIGIOUS	112.45	104.56

Table 24 - Peak accuracy by source and text type across all countries and sources (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Source / Text Type	Max Pos + Neg	Average Pos + Neg
afp		
FTXT_CLEAN	135.54	109.10
FTXT_POSNOPROPERNOUNS	134.06	108.63
FTXT_STOPWORDS	133.98	108.03
FTXT_STEMMED	133.83	108.31
FTXT_TABARIVERBS	133.51	108.07
FTXT_TABARIALLACTORS	132.71	106.83
FTXT_TONEPOS	132.05	105.43
FTXT_STEMMEDANDSTOPWORD	132.01	107.83
FTXT_TONENEG	131.96	106.39
FTXT_POSVERBS	131.57	108.25
FTXT_TONE	131.35	106.66
FTXT_POSADJS	130.61	107.46
FTXT_POSNONOUNS	130.13	108.73
META_ORG	129.11	104.49
META_GEO	128.89	105.71
META_NAME	123.81	105.76
FTXT_ETHNIC	122.30	105.42
FTXT_ETHNICRELIGIOUS	118.53	105.01
FTXT_RELIGIOUS	112.45	102.69
apmain		
FTXT_STEMMEDANDSTOPWORD	118.08	100.68
FTXT_TONENEG	115.82	100.12
FTXT_STOPWORDS	114.95	101.67
FTXT_POSVERBS	114.71	103.73
FTXT_POSADJS	114.60	100.73
FTXT_POSNOPROPERNOUNS	113.40	101.31
FTXT_TABARIVERBS	113.26	100.64
FTXT_TONEPOS	112.61	101.34
META_NAME	112.57	100.92
FTXT_TABARIALLACTORS	112.40	101.52
FTXT_STEMMED	112.16	102.38
FTXT_POSNONOUNS	112.14	101.11
FTXT_CLEAN	111.92	101.67
FTXT_TONE	111.08	100.91
FTXT_ETHNICRELIGIOUS	109.22	100.65
META_GEO	108.89	100.24
FTXT_ETHNIC	108.68	100.81
FTXT_RELIGIOUS	106.31	99.00

Table 24 (cont.)

Source / Text Type	Max Pos + Neg	Average Pos + Neg
META_ORG	100.00	100.00
xinhua		
META_GEO	127.48	103.10
FTXT_POSADJS	125.49	105.26
FTXT_STEMMED	124.28	104.75
FTXT_TONENEG	123.04	103.10
FTXT_CLEAN	122.46	104.51
FTXT_POSNONOUNS	122.10	106.25
FTXT_POSNOPROPERNOUNS	121.78	104.48
FTXT_STOPWORDS	121.36	103.74
FTXT_POSVERBS	121.19	105.25
FTXT_TABARIVERBS	121.00	104.49
FTXT_STEMMEDANDSTOPWORD	119.63	103.48
FTXT_TABARIALLACTORS	118.41	103.32
FTXT_ETHNIC	118.38	103.09
FTXT_TONE	117.83	102.61
FTXT_TONEPOS	116.42	101.83
META_NAME	116.21	103.53
META_ORG	116.11	102.87
FTXT_ETHNICRELIGIOUS	113.07	102.45
FTXT_RELIGIOUS	108.87	98.40

Table 25 - Peak accuracy by source across all text types, countries and sources (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Source	Max Pos + Neg	Average Pos + Neg
afp	135.54	106.85
xinhua	127.48	103.71
apmain	118.08	101.14

Table 26 shows the same results, but broken down by country and source across all text types. Agence France Presse yields the most accurate forecasts for all five countries, followed by Xinhua for all but Germany and Indonesia. There appears to be no correlation between each source's number of articles or events about a country and its relative accuracy at forecasting events in that country. Average article length also does not seem to be a factor, as Associated Press articles have the longest average length,

followed by Agence France Presse and Xinhua, and all three have nearly identical average words per event. Even China's long-standing censorship of labor unrest did not prevent Xinhua from covering many incidents of labor unrest in Germany ("Workers", 2010), suggesting the difference is also not driven by censorship of specific topics. One possible explanation for the Associated Press' boost in Germany and Indonesia could be the Associated Press' heavy emphasis on those two countries. Wu (2006) found that the strength of a news source's physical presence in a country is a strong predictor of the level of attention and detail it will pay that country. The Associated Press' Indonesian bureau, located in Jakarta, had by 2008 "become a showcase for [its] expanded global reach" (Ricchiardi, 2008), while its Berlin bureau oversaw coverage for all of Central Europe and was the home of its dedicated German-language newswire ("Moore", 2008).

Delving further using Germany as an example, Xinhua's coverage appears to contain more inflammatory language in some cases. For example, the Associated Press' clinically mentioned that "Berlin police are offering (EURO)5,000 ($7,180) to anyone able to help them find vandals who set fire to 18 cars parked in residential neighborhoods overnight" ("Vandals", 2011), while Xinhua described the same episode as "car arson lunacy," an "unbridled arson attack…surged up with lunacy," and quoted a local official as describing them as "a prelude to terrorism" ("Lunacy", 2011). The significant predictive power of location (META_GEO) in Xinhua is reflected in its listing of the actual districts of Charlottenburg, Tiergarten, Neu-Hohenschonhausen, and Teltow-Flaming where the arson occurred ("Lunacy", 2011), while the Associated Press noted only that the attacks occurred in "residential neighborhoods" ("Vandals", 2011). Both Xinhua and Associated Press coverage of the arson attacks quoted major politicians, but framed the attacks as unknown criminal activity. In stark contrast, Agence France Presse coverage of the attacks strongly framed them in a political context, focusing on the potential linkage to "left-wing extremists" with political agendas ("Wave", 2011) and the potential that cuts to police

staffing instituted by one of the major political parties had caused the deterioration in security ("Consecutive", 2011). In essence, while the first two sources simply reported the factual details of the arsons, Agence France Presse discussed their political context and potential impact, placing them within a strong political framework.

This same pattern is seen with another major event, the resignation of the German President over controversial remarks he made about the military. Xinhua and Associated Press once again reported the factual news of his resignation and its immediate context ("Koehler", 2010; "Remarks", 2010), while Agence France Presse contextualized the news in terms of its broader political impact on Germany's government ("Twin", 2010). While all three sources cover German politics in detail, Agence France Presse stands out for framing its coverage in terms of its political impact. In fact, FTXT_TABARIALLACTORS, which reflects mentions of global political leaders and organizations, is nearly twice as predictive for Agence France Presse coverage across all five countries as it is for Xinhua or the Associated Press. It is also suggestive that the top four most-accurate text types for Agence France Presse are all versions of the original raw text, rather than the various theoretically-filtered versions, as this is further evidence that the specific language lending its accuracy is not found in the other dictionaries. In contrast, Xinhua and Associated Press rely more heavily on the filtered surrogates like location mentions and emotional language.

Table 26 - Peak accuracy by country and source across all text types (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Country / Source	Max Pos + Neg	Average Pos + Neg
brazil		
afp	127.80	103.51
xinhua	127.48	105.60
apmain	114.60	100.69
egypt		
afp	135.54	115.70
xinhua	125.49	106.81
apmain	114.71	102.24
germany		
afp	125.99	105.82
apmain	115.82	100.58
xinhua	112.93	101.16
indonesia		
afp	120.27	102.90
apmain	118.08	101.22
xinhua	113.82	101.76
south africa		
afp	120.22	103.77
xinhua	117.32	100.63
apmain	108.01	99.93

Table 27 repeats the results of Table 9, but looking across all five countries and all three sources. Ranking the event Quad Classes by their average accuracy yields an ordering identical to that of Table 9, while ranking them by peak accuracy across all countries and sources still places Verbal Cooperation as the most readily forecasted, but ranks Material Cooperation as the next-most accurate, with Verbal Conflict and Material Conflict the fourth and fifth, respectively. Table 28 breaks these results down by country, illustrating that Verbal Cooperation is indeed the most predictable event class across all five countries. The ordering of the remaining categories varies significantly from country to country, however Table 29 shows it does not appear to be correlated with the number or relative distribution of each event class in the training or testing periods. Thus, the next section will focus on teasing out the underlying patterns in coverage of each country.

Table 27 - Peak accuracy by event type across all countries and sources (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Event Quad Class	Max Pos + Neg	Average Pos + Neg
quadverbcoop	134.06	106.24
allcount	135.54	105.86
quadmatconf	121.33	101.87
quadverbconf	123.51	101.62
quadmatcoop	125.67	100.67

Table 28 - Peak accuracy by country and event class across all text types (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Country / Event Class	Max Pos + Neg	Average Pos + Neg
brazil		
quadverbcoop	127.80	105.59
allcount	121.47	105.07
quadverbconf	115.05	99.59
quadmatconf	112.13	100.77
quadmatcoop	109.61	100.05
egypt		
allcount	135.54	110.41
quadverbcoop	134.06	111.35
quadmatcoop	125.67	103.30
quadverbconf	123.51	105.79
quadmatconf	121.33	105.12
germany		
quadverbcoop	125.99	104.68
allcount	122.44	104.38
quadverbconf	118.08	100.81
quadmatconf	115.82	102.08
quadmatcoop	112.56	99.73
indonesia		
quadverbcoop	120.27	103.68
allcount	116.21	103.29
quadmatconf	109.54	99.83
quadverbconf	107.79	99.16
quadmatcoop	106.37	99.77
south_africa		
allcount	120.22	104.16
quadverbcoop	119.15	101.30
quadverbconf	105.56	100.06

Table 28 (cont.)

Country / Event Class	Max Pos + Neg	Average Pos + Neg
quadmatcoop	105.48	100.09
quadmatconf	104.17	99.98

Table 29 – Percent of all events in combined training and testing periods for each country by Quad Class

Country	%VerbCoop	%MatConf	%MatCoop	%VerbConf
brazil	70.96	9.77	10.04	9.24
egypt	64.51	14.07	9.61	11.81
germany	66.44	13.01	10.05	10.51
indonesia	58.87	20.27	12.18	8.68
southafrica	66.85	13.90	9.99	9.27

5.3.2 PEERING INTO THE MODELS: UNDERSTANDING COUNTRY-LEVEL DRIVING FACTORS

As the tables above demonstrated, there is little conformity among the five countries or three sources regarding the textual indicators and event types most readily forecasted. While Agence France Presse is the most predictive source overall and Verbal Cooperation events the most predictable, the evidence supporting why this might be is scarce. Thus, this section seeks to tease apart the models in more detail to understand some of the underlying patterns they are learning. To begin with, Table 30 shows the most accurate text type across the three sources for each country. As with the earlier results, each country has a ranking slightly different than the others, with few larger-scale patterns readily identifiable amongst them.

Table 30 - Peak accuracy by country and text type across all event classes (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Country / Text Source	Max Pos + Neg	Average Pos + Neg
brazil		
FTXT_POSNOPROPERNOUNS	127.80	104.79
META_GEO	127.48	104.05
FTXT_CLEAN	124.06	105.04

Table 30 (cont.)

Country / Text Source	Max Pos + Neg	Average Pos + Neg
FTXT_STEMMED	123.92	104.93
FTXT_STEMMEDANDSTOPWORD	120.80	103.53
FTXT_POSNONOUNS	120.47	106.31
FTXT_POSVERBS	120.01	105.48
FTXT_TABARIVERBS	119.57	104.12
FTXT_TABARIALLACTORS	119.49	103.29
FTXT_POSADJS	118.24	104.41
FTXT_STOPWORDS	117.07	103.53
FTXT_TONE	116.86	102.58
FTXT_ETHNIC	113.07	103.43
FTXT_ETHNICRELIGIOUS	113.07	103.19
FTXT_TONEPOS	112.79	103.26
META_NAME	111.48	103.76
FTXT_TONENEG	110.91	101.84
META_ORG	109.61	100.93
FTXT_RELIGIOUS	105.31	100.27
egypt		
FTXT_CLEAN	135.54	110.40
FTXT_POSNOPROPERNOUNS	134.06	110.62
FTXT_STOPWORDS	133.98	109.96
FTXT_STEMMED	133.83	111.13
FTXT_TABARIVERBS	133.51	110.84
FTXT_TABARIALLACTORS	132.71	109.37
FTXT_TONEPOS	132.05	107.95
FTXT_STEMMEDANDSTOPWORD	132.01	109.82
FTXT_TONENEG	131.96	109.15
FTXT_POSVERBS	131.57	112.23
FTXT_TONE	131.35	109.27
FTXT_POSADJS	130.61	111.49
FTXT_POSNONOUNS	130.13	111.04
META_ORG	129.11	111.03
META_GEO	128.89	107.06
META_NAME	123.81	107.24
FTXT_ETHNIC	122.30	105.74
FTXT_ETHNICRELIGIOUS	118.53	104.44
FTXT_RELIGIOUS	112.45	101.66
germany		
FTXT_POSVERBS	125.99	104.34
FTXT_CLEAN	124.61	104.59
FTXT_TABARIVERBS	123.95	103.56
FTXT_STEMMED	123.84	104.38
FTXT_POSADJS	123.23	102.98

Table 30 (cont.)

Country / Text Source	Max Pos + Neg	Average Pos + Neg
FTXT_POSNOPROPERNOUNS	123.05	103.91
FTXT_STOPWORDS	122.00	104.44
FTXT_POSNONOUNS	121.47	104.13
FTXT_STEMMEDANDSTOPWORD	121.45	103.12
FTXT_TABARIALLACTORS	121.44	102.73
META_GEO	121.16	102.58
META_NAME	118.70	103.29
FTXT_TONE	118.26	103.02
FTXT_ETHNICRELIGIOUS	118.08	103.09
FTXT_TONENEG	117.68	103.06
FTXT_TONEPOS	117.15	101.63
FTXT_ETHNIC	116.30	103.10
META_ORG	115.12	102.02
FTXT_RELIGIOUS	110.14	101.29
indonesia		
FTXT_STEMMEDANDSTOPWORD	120.27	102.91
FTXT_TABARIALLACTORS	120.09	102.91
FTXT_POSNONOUNS	119.30	103.01
FTXT_CLEAN	117.03	103.36
FTXT_TONE	116.21	101.39
FTXT_ETHNIC	116.06	102.28
FTXT_STEMMED	115.71	103.32
FTXT_STOPWORDS	115.15	102.76
FTXT_ETHNICRELIGIOUS	114.71	102.07
FTXT_POSVERBS	114.28	103.12
FTXT_TONEPOS	113.92	101.31
FTXT_POSNOPROPERNOUNS	113.73	103.37
FTXT_TABARIVERBS	113.72	102.43
FTXT_POSADJS	113.31	101.71
META_GEO	112.38	101.26
FTXT_TONENEG	111.32	101.34
META_NAME	106.99	101.59
META_ORG	106.64	100.09
FTXT_RELIGIOUS	106.41	99.79
south_africa		
FTXT_STEMMEDANDSTOPWORD	120.22	101.68
FTXT_CLEAN	119.15	103.83
FTXT_STOPWORDS	117.94	102.12
FTXT_POSNOPROPERNOUNS	117.70	102.54
FTXT_ETHNIC	117.36	102.10
FTXT_POSNONOUNS	117.09	104.62
FTXT_ETHNICRELIGIOUS	115.88	101.97

Table 30 (cont.)

Country / Text Source	Max Pos + Neg	Average Pos + Neg
FTXT_POSADJS	115.70	102.82
META_GEO	115.65	100.91
FTXT_TONENEG	114.97	100.98
FTXT_POSVERBS	114.57	103.02
FTXT_TABARIALLACTORS	114.24	101.34
FTXT_TABARIVERBS	113.91	102.03
FTXT_TONE	113.60	100.63
FTXT_STEMMED	112.95	101.84
META_NAME	110.83	103.29
FTXT_TONEPOS	107.84	99.49
META_ORG	106.64	102.24
FTXT_RELIGIOUS	105.65	100.66

Looking to Germany once again, it is the only country for which META_NAME approaches at least the middle of the accuracy ranking, in this case using Agence France Presse source text and predicting all event types with a High Event threshold of 7 events per day. Table 31 lists the top 10 of the 36 terms in this model most predictive of future High Event days. Upon close examination, each entry is a major European leader or leader of Iran or Afghanistan. The last name, John Demjanjuk, is a Ukrainian-American who was accused of being a prison guard at a Nazi extermination camp and during the period of analysis underwent a series of high-profile court decisions. In essence, the model has learned that when foreign leaders are discussed in context with Germany in Agence France Presse news coverage, it indicates a likely impending summit with resulting cooperative agreements. In particular it has learned that meetings with Britain's Gordon Brown or France's Nicolas Sarkozy are the most likely to result in a subsequent summit and action.

Table 31 – Most predictive terms of future events for Agence France Presse coverage of Germany for META_NAME (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Term	Low Mean	Low Var	High Mean	High Var	High Bias
gordon_brown	0.76	3.77	3.72	9.80	2.96
nicolas_sarkozy	1.09	4.39	3.41	8.66	2.33
dmitry_medvedev	0.40	3.07	2.21	9.07	1.81
george_w_bush	0.32	3.50	2.13	8.27	1.80
mahmoud_ahmadinejad	1.55	6.76	3.20	10.27	1.64
angela_merkel	3.72	6.62	5.31	7.86	1.59
barack_obama	4.04	6.70	5.27	7.25	1.23
david_miliband	0.00	0.00	0.97	5.98	0.97
hamid_karzai	0.42	4.55	1.08	6.72	0.66
john_demjanjuk	0.25	2.76	0.87	6.87	0.61

Repeating this for Xinhua's coverage of Germany as viewed through META_NAME, the peak accuracy is just 2.5% better than random chance, with just 11 terms total in the model, seen in Table 32. Once again, the names are all European leaders, with the highest name being the former President of Poland and a significantly increased emphasis on Russia, including Dmitry Medvedev, Vladimir Putin, and Sergei Ivanov (a high-ranking Russian defense minister), as well as Middle Eastern leaders. Repeating for META_NAME for Associated Press, just two names are captured: Angela Merkel and Barack Obama, linking the country solely to the United States. Three different sources reporting on the same country over the same time period thus reflect three very different geographic and political lenses through which that country is viewed. This is seen in their respective descriptions of a meeting between UK Prime Minister Gordon Brown and German Chancellor Angela Merkel in April 2010. Agence France Presse mentioned that the meeting followed discussions earlier in the week with US President Barack Obama and French President Nicolas Sarkozy regarding Iran ("Tougher", 2010), while Xinhua mentioned only that the two met ("Meet", 2010), and the Associated Press mentioned only that "Senior diplomats from Britain, the U.S., France, Germany, Russia and China" were part of the discussion, but did not mention them by name ("Back", 2010).

Table 32 - Most predictive terms of future events for Xinhua coverage of Germany for META_NAME (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Term	Low Mean	Low Var	High Mean	High Var	High Bias
lech_kaczynski	0.00	0.00	3.92	17.58	3.92
barack_obama	1.66	10.92	5.46	15.12	3.80
angela_merkel	4.09	10.43	6.78	11.06	2.69
gordon_brown	0.51	4.03	2.14	9.05	1.62
dmitry_medvedev	1.24	8.14	2.33	8.58	1.09
sergei_ivanov	0.38	3.66	1.29	6.72	0.91
ali_larijani	0.41	3.97	1.29	6.72	0.88
benjamin_netanyahu	0.94	5.22	1.73	7.40	0.79
anders_fogh_rasmussen	0.47	4.61	0.92	6.84	0.45
nicolas_sarkozy	1.89	10.30	2.13	7.74	0.24
vladimir_putin	0.47	4.61	0.69	5.13	0.22

Table 33 displays all ten terms from the META_NAME model for Agence France Presse coverage of Brazil, reflecting its strong focus on Latin American leaders. The presidents of the United States and France and the US Secretary of State are the only non-regional names to be captured in the model. Much as Egypt's model captured its prominent role in mediating regional Middle East conflicts and Germany's captured its Euro-centric position, Brazil's entries reflect its more regional diplomatic role. Table 34 explores this further, displaying the term list from the META_GEO field of Xinhua's Brazilian coverage, which peaks at 27.48% better than random chance. Here Honduras is the most predictive locative name, reflecting its 2009 coup that caused ripples of discussion across Latin America and prompted another flurry of coverage in 2011 when the country was readmitted to the Organization of American States and Brazil reestablished diplomatic relations with it ("Appoint", 2011).

Table 33 - Most predictive terms of future events for AFP coverage of Brazil for META_NAME (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Term	Low Mean	Low Var	High Mean	High Var	High Bias
luiz_inacio_lula	1.63	6.03	7.45	15.92	5.82
manuel_zelaya	0.89	6.03	2.85	12.83	1.96
nicolas_sarkozy	0.38	3.31	2.16	9.94	1.79
hillary_clinton	0.57	4.45	2.19	9.87	1.62
roberto_micheletti	1.05	6.65	2.54	12.97	1.50
barack_obama	1.26	6.20	2.72	8.71	1.46
mahmoud_ahmadinejad	0.84	6.59	2.13	12.42	1.28
jose_serra	0.55	6.81	1.24	8.47	0.70
oscar_arias	0.35	3.65	0.84	6.44	0.49
hugo_chavez	1.14	5.91	1.60	6.64	0.46

Table 34 - Most predictive terms of future events for Xinhua coverage of Brazil for META_GEO (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Term	Low Mean	Low Var	High Mean	High Var	High Bias
honduras	13.59	29.00	33.02	56.34	19.44
haiti	13.26	70.61	29.06	128.43	15.79
france	16.06	50.78	31.39	103.60	15.33
costa	7.37	16.55	17.70	32.86	10.33
rica	6.78	15.91	16.93	31.65	10.14
portugal	2.76	9.39	12.32	49.65	9.56
spain	5.17	15.10	13.98	31.18	8.82
chile	8.05	14.92	15.84	24.21	7.80
venezuela	10.33	19.41	18.10	27.12	7.78
bolivia	7.28	16.69	14.74	24.52	7.46

It is clear that when it comes to political leaders and geography, the most predictive terms are highly localized and strongly steeped in local political and cultural context. Even within a single country and text type, each news agency covers global events from a different perspective, such as biasing towards Eastern or Western political leaders or focusing on the domestic political implications of major events. Continuing the exploration of Brazil, Table 35 looks more broadly at the language predictive of future events, showing the top 10 of the 522 terms appearing in the FTXT_CLEAN model (its third most

accurate) for Agence France Presse coverage of the nation. A manual review of coverage during the training period reveals that the bias towards aviation-related terms and France is largely driven by the Air France Flight 447 Brazil-to-France crash in the Atlantic Ocean in June 2009. This crash resulted in a regular stream of news coverage and cooperative agreements regarding the subsequent search and investigation ("Year and a half", 2009). Simultaneously, Brazil is home to Embraer, "the third-biggest commercial aviation manufacturer in the world whose success has made it one of the symbols of Brazil's economic boom", and whose planes are often purchased through large government-based cooperative agreements in countries like China, that result in significant coverage ("Embraer", 2010).

Table 35 - Most predictive terms of future events for Agence France Presse coverage of Brazil for FTXT_CLEAN (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Term	Low Mean	Low Var	High Mean	High Var	High Bias
air	5.45	26.91	23.84	94.35	18.39
iran	6.47	23.80	24.76	74.04	18.29
turkey	2.16	12.12	17.69	82.43	15.54
nuclear	3.81	15.85	17.56	55.24	13.75
france	5.36	22.01	18.40	65.35	13.05
deal	2.81	10.51	13.17	46.16	10.36
atlantic	2.35	15.77	12.37	50.53	10.02
plane	1.73	9.14	10.81	44.81	9.08
tehran	1.63	9.32	9.87	38.76	8.24
jet	2.10	11.67	10.00	36.28	7.91

Turning to Indonesia, Table 36 shows the top 10 of the 322 terms in the Agence France Presse FTXT_CLEAN model, which demonstrate a strong centralization on the repeated cycle of earthquakes and tsunami watches that plague the nation, including a 2004 tsunami in which 168,000 were killed ("Seismologists", 2010). Here the model has learned that each time there is a report of an earthquake, there is a subsequent report of a tsunami, which in turn is immediately followed by a massive influx of foreign aid and countless verbal statements of solidarity or condolences from foreign leaders ("Foreign

aid", 2009), as well as invariably accusations of corruption and mismanagement of disaster response activities ("Failures", 2010). Indeed, this is very similar to the kinds of rulesets learned by Radinsky & Horvitz (2013) regarding the sequencing of droughts, storms, and cholera outbreaks. In addition, US President Obama's childhood ties to the country led to its being the first Muslim-majority country visited by Secretary of State Hilary Clinton and she was involved in a number of diplomatic exchanges with the country over this period ("Clinton", 2009).

Table 36 - Most predictive terms of future events for Agence France Presse coverage of Indonesia for FTXT_CLEAN (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Term	Low Mean	Low Var	High Mean	High Var	High Bias
clinton	1.52	12.98	7.20	32.33	5.68
quake	3.46	12.17	8.39	28.30	4.93
tsunami	2.90	16.25	6.97	37.05	4.06
sumatra	1.45	9.13	5.42	24.78	3.97
earthquake	1.36	7.28	5.10	22.97	3.74
least	1.20	10.57	4.93	23.45	3.73
struck	1.09	5.87	4.34	16.91	3.26
killed	3.01	10.07	6.14	18.51	3.14
magnitude	1.44	7.12	4.49	16.60	3.04
padang	0.77	11.52	3.80	24.82	3.03

Continuing the analysis of Brazil and Indonesia through the eyes of Agence France Presse, but narrowing the focus to Material Conflict events, rather than all events, the following two tables outline the top terms for their respective FTXT_CLEAN models. As with the Verbal Cooperation-dominated All Events model, the top terms in Table 37 revolve around the Air France crash, which appears to have dominated Agence France Presse coverage during this time period. Also prominent are terms related to the 2010 Haitian earthquake, which killed Brazilian military personnel stationed there ("Boost", 2010) and damaged several of the country's facilities. The death of Brazilian military personnel in particular was tied in the Agence France Presse narrative to the country's desire to "recast [itself] as a country able to

project itself abroad" and "to become the region's pre-eminent military power" ("Boost", 2010). This in turn lead to considerable discussion about other conflicts the country might become involved in, including Lebanon, and foreign conflicts it was currently involved in, such as leading the 2004 UN stabilization mission in Haiti ("Slow return", 2010). Thus, once again, the news agency contextualized its coverage in political terms, here capturing the political dimension of natural disaster.

Table 38 shows the same results for Indonesia. While the focus on tsunami-related language of Table 36 is still seen, here it appears to be driven by widespread protests over the government's handling of relief funds. A 2009 summit that partially focused on earthquake relief assistance was marred by thousands of protesters marching across the country against the president "in response to a series of scandals which have damaged the credibility of [his] government" leading the nation be called "one of the most corrupt countries in the world" ("Bali", 2009). Even a subsequent visit by US President Obama a month after a tsunami killed 430 and a volcano displaced 100,000 led thousands of protesters to take to the street holding arguing "with the disasters happening in Indonesia, his visit will only add to our grief" ("Islamists", 2010). On top of all of this was strong coverage of an increase in suicide attacks targeting Western tourists and hotels in the country during the period ("Vigilance", 2010).

Table 37 - Most predictive terms of future Material Conflict events for Agence France Presse coverage of Brazil for FTXT_CLEAN (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Term	Low Mean	Low Var	High Mean	High Var	High Bias
air	3.96	24.10	26.55	96.33	22.59
france	3.86	19.22	21.53	72.27	17.67
haiti	2.49	19.32	17.85	62.71	15.36
atlantic	1.53	14.84	14.90	58.93	13.38
jet	1.44	10.67	11.92	44.80	10.48
flight	1.39	11.27	10.07	44.79	8.68
plane	2.03	12.17	10.44	47.83	8.41
quake	1.78	12.64	9.92	33.32	8.14
missing	0.64	5.43	8.16	45.16	7.52
paris	0.93	6.91	8.38	33.74	7.45

Table 38 - Most predictive terms of future Material Conflict events for Agence France Presse coverage of Indonesia for FTXT_CLEAN (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

	Low Mean	Low Var	High Mean	High Var	High Bias
tsunami	2.04	10.09	10.58	45.26	8.54
killed	2.37	7.87	8.82	22.14	6.45
people	4.03	8.69	9.47	19.27	5.44
least	0.98	7.17	6.16	28.49	5.18
volcano	0.59	5.21	5.45	30.32	4.86
islands	0.18	2.44	4.64	21.95	4.46
hotels	0.41	5.50	4.83	32.86	4.41
missing	1.63	8.26	5.54	23.32	3.90
powerful	0.21	3.10	3.97	17.93	3.76
injured	0.28	3.04	3.96	22.04	3.67

5.3.3 TESTING CROSS-COUNTRY MODELS

As the results thus far have demonstrated, each of the models appears to be learning highly source- and country-specific features that are not likely to be generalizable across countries. For example, learning that discussions of earthquakes are often followed by tsunamis and subsequent diplomatic agreements and protests in Indonesia is not likely to offer much predictive power over estimating future events in

Brazil. Likewise, learning the political leaders and nations must closely associated with Brazil's ascendency to the global stage as a preeminent Latin American diplomatic power is not likely to be helpful in forecasting Indonesian events.

To test this quantitatively, the Agence France Presse FTXT_CLEAN models for forecasting Verbal Cooperation events in Indonesia and Brazil were selected. To compute the baseline accuracy of forecasting events in the same country it was trained on, the model was first trained and tested on the FTXT_CLEAN version of Agence France Presse coverage of Indonesia. Here, the model was given news coverage of Indonesia and Indonesian Verbal Cooperation events during the training period and again during the testing period. This yielded a baseline of 17.03% better than random chance, and was repeated for Brazil, yielding 24.06% over random chance. This demonstrates the accuracy of the model when trained and tested on the same country. In the second configuration, the model was trained on Indonesian news coverage and Indonesian Verbal Cooperation events, but then tested on Brazilian news coverage and used to forecast Brazilian Verbal Cooperation events. The resulting accuracy indicates the model performed no better than random chance. Similarly, when the Brazilian-trained model was applied to forecasting Indonesian events using Indonesian coverage, the accuracy plummeted below the baseline, to just 5.72% above random chance.

Thus, as evidence from examining each of the forecasting models in detail has suggested, the models, while predictive of future events, gain their predictive power by learning very narrowly constrained linguistic indicators descriptive of metanarratives distinct to a particular cultural environment. Indeed, much as a new human analyst learns that the political processes which drive events in Indonesia are highly distinct from those in Brazil, the models here have indirectly learned those same discerning rules.

In fact, the improper application of general patterns from one country to forecast those in another has been at the root of many of the failures of past forecasting work (O'Brien, 2010).

Table 39 – Testing Indonesian and Brazilian Agence France Presse models on cross-country forecasting for Verbal Cooperation events using FTXT_CLEAN (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Source	Predicting	True Pos	True Neg	Pos + Neg
Indonesia/AFP	Indonesia/quadverbcoop	34.19	82.85	117.03
Indonesia/AFP	Brazil/quadverbcoop	82.20	17.37	99.57
Brazil/AFP	Brazil/quadverbcoop	52.63	71.43	124.06
Brazil/AFP	Indonesia/quadverbcoop	94.02	11.70	105.72

5.3.4 REVISITING PROTESTS

The results of Table 10 earlier suggested that forecasting specific event types was considerably less accurate than forecasting broad classes of events. However, it was based on Xinhua coverage of Egypt and used a 6-year training period to forecast a 6-year testing period. Using the narrower 2-year training and 2-year testing period of above and looking across all five countries and all three sources, Brazil and Egypt are the only two countries to achieve greater than 5% better than random chance at forecasting Protest events. In both cases, Agence France Presse provided the most accurate coverage, with Egypt reaching 15.51% better than random chance (90.07% True Positive / 25.44% True Negative) and Brazil reaching 6.65% (100% True Positive / 6.65% True Negative). In the case of Egypt, it was location mentions via META_ORG that yielded the most predictive results, shown in Table 40. The model appears to have learned that discussions of the United Nations, the UN Security Council, the Arab League, and the European Union all are suggestive of future protests, while mentions of the World Health Organization and International Criminal Court are less suggestive of future protests. Indeed, it appears the ICC mentions are related to Sudanese President Omar al-Beshir's visit to Egypt during this

period, which yielded substantial coverage tied into discussions of United Nations activity in Darfur ("Beshir", 2009). The discussion of the World Health Organization, on the other hand, appears to be centered on a partnership with Egypt to reduce tobacco use ("Alexandria", 2010) and regular updates on the spread of Avian Flu through the region ("Bird flu", 2010).

Table 40 - Most predictive terms of future Protest (Code 1400) events for Agence France Presse coverage of Egypt for META_ORG (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Term	Low Mean	Low Var	High Mean	High Var	High Bias
un_security_council	2.03	13.33	11.01	29.54	8.98
united_nations	1.13	7.60	9.48	26.31	8.35
arab_league	0.57	4.60	6.22	14.87	5.65
european_union	1.59	7.25	6.28	12.82	4.69
world_health_organization	0.66	6.10	0.00	0.00	-0.66
international_criminal_court	0.75	5.91	0.00	0.00	-0.75

For Brazil, only META_GEO has measurable predictive insight. The strongest indicator is Honduras (and Tegucigalpa its capital), which is a result of the Brazilian embassy providing sanctuary to deposed Honduran President Manuel Zelaya, leading to wide-spread protests against Brazil in Honduras ("Zelaya", 2009). His replacement was forced to withdraw from a European-Latin American summit the following year amid further protests and boycott threats from other Latin American nations ("Boycott", 2010). The strong connection to Iran is due to Brazil's emerging role as a regional diplomatic superpower and its "friendly ties" with Iran ("Repress", 2010), while the connection with Venezuela appears driven by regular comparisons against its neighbor. For example, when Brazilian humorists took to the streets to protest a new ban on political satire during the 2010 presidential campaign, they widely compared the ban to "the example of Venezuela" ("Humorists, 2010). Of interest, while discussion of France in Agence France Presse coverage of Brazil is closely tied during this time period to Verbal Cooperation events, it is only ranked 68th out of 224 terms for indicating future protests, reflecting that

the discussion around France was linked to cooperation between the two governments, rather than protests.

Table 41 - Most predictive terms of future Protest (Code 14XX) events for AFP coverage of Brazil for META_ORG (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Term	Low Mean	Low Var	High Mean	High Var	High Bias
honduras	14.62	48.32	141.97	157.75	127.34
venezuela	8.34	12.03	67.23	9.16	58.89
iran	6.55	19.89	62.86	88.89	56.30
tegucigalpa	3.69	19.11	47.28	66.87	43.59
denmark	3.39	9.30	46.73	27.21	43.34
kyoto	0.65	5.39	39.88	56.39	39.23
belgium	1.81	7.85	40.44	57.20	38.63
tehran	1.91	13.25	37.57	53.13	35.66
israel	5.01	11.15	40.37	57.08	35.35
libya	2.39	11.34	37.64	53.23	35.25

5.3.5 TESTING WEEKLY FORECASTS

As the Egyptian revolution demonstrated, even the ability to forecast events just 24 hours in advance can sometimes be as just as critical as the ability to forecast trends months or years in advance. In keeping with this, each of the experiments thus far has used the previous two days of news coverage to forecast events the following day. Yet, the desire of policy makers for longer-term forecasts that afford greater opportunities for actionable intervention (O'Brien, 2010) continues to push the forecasting field towards longer time horizons (IARPA, 2011). To test whether the classification approach to forecasting outlined here can operate on longer time horizons, Table 42 shows the results of using weekly aggregation to forecast all classes of Egyptian events using Xinhua coverage, with a training period of November 2008 to October 2010 and a testing period of November 2010 to August 2012, as above.

Here, both training and testing periods were divided into weeks running from Sunday to the following Saturday, and all articles published during that week were combined into a single input "day" for the model. Similarly, the number of events occurring in the following week were summed and associated with that input "day." This allowed an identical technical configuration and modeling pipeline to be used, with the only change being the use of an entire week's worth of news articles to forecast the total number of events to occur anytime the following week. In a production configuration, such a system would operate each Saturday evening, collecting the previous week's coverage and issuing a forecast regarding whether the following week would be a High or Low Event week. While some text types experienced a reduction in accuracy, such as FTXT_TONENEG, which now achieves accuracy no better than random chance, other text types, like META_GEO have substantially increased accuracy. In fact, META_GEO achieves 45.02% better than random chance under this configuration, correctly recognizing nearly three-quarters of both High Event and Low Event weeks.

Table 42 – Peak accuracy by text type for Xinhua coverage of Egypt using weekly aggregation (training 11/2008 – 10/2010, testing 11/2010 – 8/2012)

Text Type	High Thres	True Pos	True Neg	Accuracy	Pos + Neg
META_GEO	28	72.73	72.29	72.34	145.02
FTXT_TABARIVERBS	60	58.33	69.57	63.83	127.90
FTXT_CLEAN	60	41.67	82.61	61.70	124.28
FTXT_STOPWORDS	60	56.25	67.39	61.70	123.64
FTXT_RELIGIOUS	34	30.00	91.67	45.74	121.67
FTXT_STEMMEDANDSTOPWORD	60	62.50	58.70	60.64	121.20
FTXT_TABARIALLACTORS	60	75.00	45.65	60.64	120.65
FTXT_POSNONOUNS	60	68.75	50.00	59.57	118.75
FTXT_ETHNICRELIGIOUS	69	50.00	66.67	59.57	116.67
FTXT_ETHNIC	69	75.00	38.89	54.26	113.89
FTXT_TONEPOS	69	72.50	40.74	54.26	113.24
FTXT_STEMMED	60	54.17	56.52	55.32	110.69
FTXT_POSNOPROPERNOUNS	60	54.17	56.52	55.32	110.69
FTXT_POSADJS	60	33.33	76.09	54.26	109.42
FTXT_POSVERBS	60	52.08	54.35	53.19	106.43
FTXT_TONE	60	56.25	50.00	53.19	106.25

Table 42 (cont.)

Text Type	High Thres	True Pos	True Neg	Accuracy	Pos + Neg
META_ORG	37	4.41	100.00	30.85	104.41
META_NAME	69	95.00	9.26	45.74	104.26
FTXT_TONENEG	60	62.50	34.78	48.94	97.28

CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation has presented a novel framework enabling the high-resolution quantitative study of the interplay of the latent and physical worlds. Using this framework, it has successfully demonstrated the application of a physical event database from the political science literature, using software to automatically transform a collection of more than 4.7 million articles and 1.3 billion words into nearly 29 million global events in 310 distinct categories coded to the specific day and location of the event. This template was then applied to test the hypothesis that forecasting future physical behavior could be treated as a classification of discourse problem, using a standard Naïve Bayesian document classification model to forecast future behavior based on latent linguistic patterns.

6.1 A TEMPLATE FOR TESTING THE LATENT-PHYSICAL LINK

One of the greatest limiting factors of the existing literature on using latent indicators to forecast future physical behavior has been the lack of a framework for testing high-resolution forecasts. Leetaru (2011) could study only whole-country collapse, while Chadefaux (2012) was limited to large-bore militarized disputes and Radinsky & Horvitz (2013) looked primarily at significant mortality events. While there are many "event databases" available today that capture human behavior in the quantitative form amenable to statistical modeling, they tend to have a limited number of event categories, low spatial or temporal resolution, or are unavailable for academic research. In particular, the US Department of Defense ICEWS project offers an ideal test bed for evaluating latent forecasting measures, combining an archive of news content with extracted event records in over 300 categories. However, other than a few

small extracts created during the evaluation phase of the project, the data is unavailable for open academic research. Other initiatives such as the recent IARPA Open Source Indicators program are similarly investing considerable resources in the construction of high-resolution event datasets for the evaluation of forecasting models of interest to US national security, but again restricting them from open academic research. No currently available academic datasets offer coverage of Latin America, while the only available African database offers just 50,000 events across more than 50 countries over a decade, limiting some countries to just a few events per year.

This dissertation has therefore demonstrated the use of automated software-based coding systems from the political science literature to construct a high-resolution spatial-temporal event database for latent forecasting, in this case capturing more than 29 million global events covering all countries over 30 years. More importantly, it has shown that constructing such a database does not require an investment of millions of dollars and a team of hundreds of highly trained human coders working over the better part of a decade. Rather, a single graduate student using just a laptop computer can now rapidly construct such a wide-ranging database within the scope of a doctoral dissertation, creating a customized database covering the time periods, geographic regions, and events of interest. Even more powerfully, the software and event taxonomy used is the same that currently underpins the US military's operational global watch board that is used to monitor emerging developments across the world each day. It therefore carries with it an extensive history of validation and continuous development and means that resulting forecasting models are directly aligned with the event categories of greatest interest to the US Government.

Moving beyond coarse long-duration episodes of whole country collapse and full-scale military conflict towards capturing the occurrence of individual peaceful protests and verbal actions such as peace

appeals and aid provisioning is shown to enable an entirely new resolution of forecasting. While the current output of the TABARI system contains only limited detail about each event, the results of this dissertation have demonstrated they are sufficient for evaluating latent forecasting approaches. In addition, the TABARI system is still under active development, introducing new ethnic and religious attribute fields just a few months ago. The ability to compile events directly from the narrative text to be analyzed eliminates source differences from forecasting error, allowing more a precise understanding of error. The flexibility of the TABARI software means it is possible to rapidly expand its event taxonomy in the future to add new classes of behavior of theoretic interest to be forecasted.

6.2 MAJOR FINDINGS

Utilizing this new framework for evaluating latent forecasting models, this dissertation has demonstrated the feasible treatment of latent forecasting as a classification problem, using a Naïve Bayesian document classification system to explore the three research questions posed in Chapter 3. It is clear that even a basic Naïve Bayesian classifier, with its assumptions of term independence, can offer potentially actionable forecasts of future physical behavior. While the most accurate forecasts appear to come from examining specific textual dimensions, such as locations or actions, it is also clear that even models based on the unfiltered raw text yield accurate forecasts. This is a valuable insight, as it means that no theoretic understanding of an event type is necessary to forecast it. One can simply analyze the entire body of text preceding occurrences of an event, rather than developing customized textual surrogates based on a theoretic construct of how and why it occurs. Indeed, from an operational standpoint it has been successfully demonstrated that an analyst could simply select a set of events of

interest from the past, hand them to a classification model, and receive fully autonomous daily or weekly forecasts of future occurrences of those events.

While the accuracies of the forecasting models presented here are relatively low, they represent completely untuned models with considerable potential for accuracy improvements. From enhanced tuning of the internal Naïve Bayesian parameters, to the use of more sophisticated classification algorithms like Support Vector Machines or Random Forest models, substantial future improvements are available. In addition, programs such as IARPA's Open Source Indicators (2011) initiative are exploring whether combining large numbers of indicators in parallel improves recognition accuracy, rather than examining a single indicator at a time as this dissertation has done. An emerging literature is also demonstrating significant gains from pooling multiple forecasting models together, combining their respective forecasts (Montgomery, Hollenbach, & Ward, 2012), much as prediction markets aim to work around the limitations of the solitary expert by pooling their mutual expertise. Finally, while the use of days and weeks as forecasting horizons may at first appear to be too short of duration to be of actionable utility, it is worth noting that there are countless situations each day where an analyst needs to know what will happen the following day (IARPA, 2011).

It might at first seem trivial to forecast certain classes of events, especially "predicting" a round of Verbal Cooperation agreements after a week of discussion leading up to them. However, in the process of creating their forecasts, these models are actually contextualizing those events into their linguistic predecessor cues and codifying into a set of probability rules the language of diplomatic discourse. In this way, they enable the quantitative study of how governments prepare their citizens for conflict and for peace, but through fully automatic algorithms rather than through the laborious hand coding of works like Hunt (1997). Thus, while the original experimental design of this dissertation was based

around testing the forecasting ability of latent textual indicators, it has demonstrated that this approach has the deeper ability to directly quantitatively model the topical traces of the metanarratives that underlie the communicative sphere.

The finding that the predictive value of information does not decrease solely with time lends critical nuance to the importance that recency and novelty play in the news cycle. The norms of "newsworthiness" examined in the literature would suggest that more recent information displaces older information specifically because of its greater ability to inform on contemporary events. Intuitively, it would certainly seem that a news story about a violent attack in Syria in early 2013 would be better contextualized by information about the broader 2011-2013 revolution, rather than information on a civil disturbance from two decades prior. Yet as Agence France Presse's propensity to contextualize stories in their historical political underpinnings demonstrates, older information is often just as important to explaining the "why" of a story and how the situation reached its present point. Rather than simply decaying with time, the predictive value of information appears to be far more driven by the degree to which it aligns with current events. In the case of Egypt, it remained relatively stable in the decade prior to the 2011 revolution, occupying a consistent role in regional politics, leading to relative stability in the predictive value of news articles from throughout that period. It is also clear that the news media, far from being an objective reporter of simple factual occurrences, offers strong contextualization of societal behavior inside source-specific frameworks. Agence France Presse, for example, appears to strongly situate events in terms of their past and future political implications, while Xinhua contextualizes in terms of space. This is simultaneously confounding, as it prevents the creation of "universal" models that generalize across inputs, and reassuring, as it exhibits the kinds of cultural contextualization theoretically understood to drive news reporting.

The modeling approach used in this study is limited to learning patterns of individual word occurrences and cannot generalize beyond those words to their more abstract semantic meanings or the way those meanings cohere into narrative. Thus, these models can access the metanarratives of a country only as instantiated in the word probabilities of the media coverage of individual occurrences of that narrative, such as discussion of a specific peace summit or attack. Much like topic modeling can offer the human analyst only a probabilistic blueprint of terms suggestive of each category that must then be woven by that human into a cohesive story describing that category (Blei & Lafferty, 2007), the probability tables of the models in this dissertation are merely starting points from which an analyst could posit a myriad possible metanarratives. This is an important consideration, in that for all their quantitative might, these models do not alleviate the need for basic humanistic qualitative inquiry. Their term lists can suggest an unexpected pattern or offer findings in support of an expected pattern, but ultimately they are merely tools that can be used to explore metanarratives along quantitative dimensions. Yet, even this relatively simplistic ability to quantitatively capture topical traces of these metanarratives enables a range of analytic methods such as clustering and model comparison techniques (Steinbach, Karypis & Kumar, 2000) that could be used to autonomously align extracted models with a predefined gallery of recognized human-defined metanarratives for a given country.

Returning to the three research questions of Chapter 3, this dissertation presents the following:

- **RESEARCH QUESTION #1. What latent signatures precede physical societal-scale behavior and manifest themselves in the media in a measurable way?** The experiments of the preceding chapter have demonstrated that text classification models are capable of surfacing a number of latent features that appear predictive of future physical behavior. There does not seem to be any one signature or set of signatures that are most predictive, but rather an array of signatures

that are closely coupled with underlying potential metanarratives of a region, time period, and culture.

- **RESEARCH QUESTION #2. Are signatures universal across geographies, or keyed to each location and culture?** It appears that signatures are highly contextualized and do not generalize across news sources or countries. Signatures do appear to generalize within a country over time as long as the underlying environment and role of that country remains stable. While the specific features change, the underlying predictive power of a model applied to a stable country does not decay strictly with time as it is trained on older information.
- **RESEARCH QUESTION #3. Are signatures universal across classes of physical behavior and intensity levels?** There is significant stratification in the ability to forecast different classes of physical behavior that is largely dependent on the degree to which those events require priming. Governments and the news media appear to prime citizens well in advance of impending Verbal Cooperation events, while conflict events are more difficult to forecast, at least using the classification approach examined here. The signatures most predictive of each class of event appear to vary considerably across news sources and countries. However, episodes with higher intensity levels, as defined by the number of events recorded in that time period, are considerably more predictable than those with smaller numbers of events.

6.3 FORECASTING AS SCIENCE AND COMPARING ACCURACY

This dissertation makes considerable use of forecasting, yet rather than focusing on forecasting as an ends itself, it is utilized as a lens through which to understand the underlying latent patterns that are

most strongly suggestive of future physical behavior. Brandt, Freeman, and Schrodt (2011) offer this observation as to the value of forecasting in modeling:

> Campbell (2000) argues that forecasting contributes to political "science"; Schrodt (2010), basing his arguments on the logical positivists Hempel and Quine, goes further to assert that unless validated by prediction, models, even those lovingly structured with elaborate formalisms and references to wizened authorities, are merely "pre-scientific."

While forecasting has been used here merely as an evaluative metric, with no emphasis on maximizing the forecasting accuracy of the resulting models, it is nevertheless useful to compare the results with the current state of the art. The TABARI and CAMEO event framework used in this dissertation are widely used in the political forecasting community and produce many of the underlying datasets. Recent forecasting work on CAMEO-coded data has turned to increasingly sophisticated modeling techniques from Markov-Switching Bayesian Vector Autoregressive models (Brandt, Freeman & Schrodt, 2011) and Latent Dirichlet Allocation models (Schrodt, 2011) to Hidden Markov models (Schrodt, 2000). However, as Brandt, Freeman & Schrodt (2011) explore at length, the lack of consistency in the accuracy metrics used to describe these models makes it exceedingly difficult to compare their results. At the very least, recent papers have touted accuracies of 10-25% better than random chance (Schrodt, 2011; D'Orazio, Yonamine & Schrodt, 2011) as positive developments in forecasting development.

Such accuracy levels would suggest that many of the models in the previous chapter are competitive with current physically-based models, yet impose fewer theoretic requirements on their use – an analyst simply selects a set of events of interest and the system forecasts the likelihood of future incidents of those events. This lends considerable ease to their operational implementation in that analysts can monitor emerging situations of interest where there is not yet enough known to develop the more detailed theoretic causative models needed for physically-based forecasting approaches. Of the few

published studies involving latent models, none has attempted to forecast low-intensity unrest, so there is no established baseline against which to compare, other to note that accuracies appear promising compared with Radinsky & Horvitz's (2003) forecasts of individual storyline occurrences.

Returning to the argument of Brandt, Freeman, and Schrodt (2011), forecasting, rather than being an ends itself, can often serve as a useful tool in determining whether patterns observed in data on societal behavior are sufficiently generalizable to be "predictive" of future behavior. In other words, of the infinite descriptive patterns that may be observed preceding a given episode of interest, which of those patterns occurs most often before similar episodes in the past and future and less often when no subsequent episode occurs? A rising criticism of traditional models of conflict is that in the absence of such forecast-based testing, models record only localized patterns of little use in understanding future conflicts (Ward et al, 2010). Forecasting as applied in this dissertation is therefore a form of contextualization, which allows those patterns most "predictive" of future behavior to be surfaced most readily from a collection. While the accuracy of these resulting models appears on par with the existing literature, their use here lies in their ability to tease apart the underlying patterns that best capture the latent underpinnings of future unrest, rather than on the purely operational pursuit of optimal prediction.

6.4 METHODOLOGICAL LIMITATIONS AND THEIR MITIGATIONS

Every analytical approach has a number of limitations that constrain its ability to fully answer the research questions being explored. This can range from the representativeness and boundaries of the source data being analyzed to the ability of the statistical methods used to adequately express and

model the source patterns being tested. This section explores some of the key limitations to the methods explored in this dissertation and the actions taken to mitigate their impact on its findings.

6.4.1 DATASET BOUNDARIES

The first and most obvious limitation is the natural boundary created by the set of available event types in the CAMEO taxonomy. While the version of CAMEO used in this dissertation contains over 300 categories, it does not at present contain any event classes relating to the spread of disease or electoral disputes, both of which have become increasingly prevalent concerns in many regions of the world (IARPA, 2011).

6.4.1.1 Mitigation

The current version of CAMEO reflects several decades of research on the categories of events most commonly discussed in the news media and thus captures the most prevalent categories of activity available for analysis. In addition, the creator of TABARI and CAMEO is already applying a number of natural language processing techniques to expand the CAMEO taxonomy for a future release (Bagozzi & Schrodt, 2012).

6.4.2 DATASET ERRORS

Even the most meticulously-created event datasets will contain a certain degree of error. False positives (event records that record an event that did not actually take place) and false negatives (a missing event that actually did occur) can both skew the picture of conflict painted by the final event data (Stoll &

Subramanian, 2006). Even the most sophisticated natural language processing systems today still incur a fair degree of error when translating freeform qualitative human-generated text into codified numeric quantitative measures.

6.4.2.1 Mitigation

King & Lowe (2003) conducted one of the most extensive comparisons of automated and human coding, using a software system called VRA that is a commercial implementation of the TABARI framework used here. To compare the two approaches, the authors had trained Harvard undergraduate students code a series of 711 texts into an event taxonomy and compared their results against automated coding of the same texts. The machine achieved 85% accuracy overall, compared with 88% for human coders, and 70% accuracy at distinguishing between specific fine-grained categories, compared with 65% accuracy for the human coders. The human coders also exhibited significant variability, ranging from 56% to 72% accuracy. The only measurable difference between human and machine coders was the machine's higher propensity for false positives. However, these were found to be non-systematic, spread randomly across categories and actors. Further, Schrodt & Yonamine (2012) note that such "coding error" is statistically insignificant in forecasting work compared with the error induced by media bias, specification error, and the intrinsic randomness of human behavior.

6.4.3 ADVERSE MEDIA EFFECTS

Media-based conflict measures may be negatively impacted by a variety of effects such as selection bias, media fatigue, and helicopter journalism. Selection bias refers to the fact that only a fraction of all global activity is eventually reported in the mainstream news media (Rosenblum, 1981). From the

standpoint of using media as a remote proxy of emotional and thematic reactions to ongoing events, the filtering effect of the media is useful in its reflection of localized socio-cultural cues (Gerbner & Marvanyi, 1977). At the same time, as a catalog of human activity, it can result in over- and under-emphasis of specific event types or attributes (McCarthy, McPhail & Smith, 1996). Similarly, helicopter journalism refers to the behavior of the media in offer little coverage to certain regions of the world and subject matters until a major incident occurs, at which point media attention may suddenly increase a thousand-fold overnight. The term draws its name from the practice of mainstream journalists to literally "helicopter" into remote regions to rapidly to report on emerging situations.

For example, international media paid comparatively little attention to Haiti until the 2010 earthquake, with some major newspapers jumping from one article a month to hundreds of articles in the subsequent months. The combination of these effects can yield a sudden massive spike in coverage of a region, followed by an exponential decay curve as coverage tails off over the following week or weeks. In addition, a long literature has suggested that the news media tends to emphasize "bad" news over "good" news (Woolley, 2000), especially regarding coverage of the third world (Rosenblum, 1981). There are also indications that the news media worldwide is becoming steadily more negative, accelerating in the Internet era with increased competition among news outlets that were previously isolated by geography (Leetaru, 2011). This could reflect the media selecting more negative stories, increased coverage of regions of the world that were previously inaccessible to the media, or a natural evolution in the use of language. In contrast, Twitter appears to emphasize positive stories, at least with respect to certain topics (Wong et al, 2012).

6.4.3.1 Mitigation

Empirical work has suggested that the incorporation of multiple media sources, especially local and regional sources, can strongly mitigate the impact of selection bias, media fatigue, and helicopter journalism, compared with relying only on a single Western international news source like the New York Times (Gerner & Shrodt, 1996). The progression of the media towards negativity is nearly linear, suggesting it will have a minimal impact on findings and can be modeled out as necessary.

6.4.4 CENSORSHIP AND MEDIA INTERFERENCE

The final primary limitation on this dissertation's results is the impact of censorship, both active and passive, on the emotional and thematic content of media that may restrict the types of signals that would portend future unrest. There are three primary areas where governments may directly influence the media sphere in ways that may skew results. The first is through censorship, the active removal of messages deemed objectionable, the second is through inorganic messaging campaigns that saturate the online discourse, and the third is through self-censorship in which the government creates an environment so hostile to dissent that discussants self-restrict their expression to avoid discussing off-limits topics. Each of these may limit the presence of early warning signals in the media environment, dampening or subverting signals that otherwise would provide strong early indicators of future unrest.

6.4.4.1 Mitigation

Censorship and governmental interference are not restricted to media channels or remote population assessment. Even field surveys and other forms of sampling are subject to this, as citizens restrict their

views to prevent retaliation and governments take action to remove prominent dissenting individuals from being able to propagate their views to the rest of society. Thus, the effects described herein do not impact media-derived latent indicators any more strongly than other conflict forecasting signals.

6.4.5 CENSORSHIP: BLOCKING MESSAGES

The outright prohibition of objectionable material through active prevention and removal strategies, known as censorship, is perhaps the oldest and most wide-spread form of interference in the media environment. The OpenNet Initiative maintains profiles on 55 countries that utilize various forms of technical or legal censorship of the Internet (OpenNet, online). China is perhaps the most noted for its wide array of technical mechanisms and long history of restricting dissenting views through the media and is continually evolving and expanding its arsenal of censorship approaches as new forms of media emerge (Hunt, 2012). Even European countries have turned to various forms of censorship during crisis: as political turmoil accelerated in Greece in 2012, the publishing of political opinion polls was legally banned (Reuters, 2012). Censorship in many countries occurs primarily through "silent" channels such as technological filtering or intimidation, but may also occur through legal channels. The online database Chilling Effects (http://www.chillingeffects.org/) maintains a list of legal requests to censor material, archiving nearly a quarter-million requests over just the last 12 months.

The increasingly global nature of electronic media has also led to a rising number of censorship conflicts stemming from culturally-sensitive material rather than anti-government commentary. For example, in May 2012, Pakistan blocked all access to Twitter from within its borders because of references to a Facebook competition to post images of Muhammad, the Muslim prophet (Sayah, 2012). Censorship demands have increased to such a point that Twitter implemented an official policy in 2012 that it

would censor tweets on a country-by-country basis so that a given country could request that a selected tweet or set of tweets be rendered inaccessible to users in those countries (CNN, 2012).

6.4.5.1 Mitigation

Censorship might at first appear to be an insurmountable obstacle to extracting early warning indicators from media coverage. Intuitively, the very nations with restrictive environments and growing levels of anti-government organizing that might be expected to be the most fragile would also be the most likely to utilize various forms of censorship to block the anti-government discourse that would offer an early warning indicator of its growth. However, motivated opposition groups will always find a way around even the most comprehensive censorship efforts. For example, the 2011 pro-democracy rallies in China became known as the "jasmine revolution" as organizers used the jasmine flower as a silent symbol of dissent. Instead of messages of "revolution" or "protests," online message boards lit up with discussions of botany and the jasmine flower (Ramzy, 2011). In Greece, when political polls were officially banned, political parties released their own confidential polls, the results of which were widely distributed through media channels (Reuters, 2012). Even in North Korea, perhaps the most censored country in the world, citizens find a myriad ways to work around those censorship controls and produce and consume restricted media (Kretchun & Kim, 2012).

Most critically, censorship efforts are not necessarily directed at anti-regime messages. In the case of China, a recent study (King, Pan, Roberts, 2012) found that censorship efforts there are directed primarily at mobilization campaigns and messaging that might promote physical action, while other forms of anti-regime commentary are largely permitted as a form of "pressure-relief valve." In this way, the Chinese government has made a decision to allow the Internet to be used as a forum for citizens to

"vent" their frustrations and largely leave this discourse unfettered, intervening only when that discourse escalates to the level of a call to action. This suggests that even in a heavily-censored nation like China, at least the early stages of anti-government rhetoric may be largely preserved intact, arguing that censorship controls may not entirely squelch latent indicators of unrest.

6.4.6 CENSORSHIP: SATURATING MESSAGES: ORGANIC VERSUS INORGANIC CAMPAIGNS

Another mechanism through which the state may negatively influence the media environment is the use of inorganic media campaigns, known as "astroturfing" (Lee, 2010). In contrast to organic campaigns in which a large number of individuals sharing a common interest discuss and promote a topic to the degree that it becomes heavily prevalent across the Internet narrative, inorganic campaigns employ very large collectives of paid users or automated software tools to pour out commentary across all available forums to "flood" the online discourse to support or squelch a particular topic. China has created perhaps the most sophisticated example of astroturfing, combining electronic censorship tools with the so-called "Internet Water Army" (Chen et al, 2011).

6.4.6.1 Mitigation

Mainstream media is more resistant to astroturfing in that its gatekeeping editorial structure and professionalized corps of reporters acts as a bulwark against such message flooding. More critically, however, while social media is amorphous with respect to national boundaries (Chinese astroturfers can saturate a discussion on the American-based Twitter platform), mainstream media tend to be subject to such censorship only with respect to the demands of their host country. While a Chinese newspaper might be prohibited from covering a labor dispute in Beijing, news outlets throughout the rest of the

world will likely not adhere to such restrictions. In the case of the 2011 Egyptian revolution, Leetaru (2011) found that while the tone of Egyptian media nearly flat-lined as the country neared collapse, reflecting a rapidly-increasing level of government censorship and intervention, this was more than compensated for by using a basket of media sources from across the world. Thus, the use of multiple mainstream media sources, including local and regional sources, should minimize the impact of such inorganic campaigns.

6.4.7 CENSORSHIP: SELF-CENSORSHIP: HARASSMENT OF THE MEDIA

The third category of governmental influence of the media is perhaps the hardest to measure. Here the government chooses not to directly manipulate the media sphere, but rather uses punitive mechanisms to discourage the dissemination of objectionable messages. For example, bloggers or journalists who post "subversive" content or who are suspected of facilitating information dissemination may be arrested or detained for lengthy periods of time to discourage others (Gladstone & Afkhami, 2012). Over time this leads to an environment of fear where individuals self-censor and avoid posting material that might have even the possibility of being construed as prohibited to avoid a similar fate. This would dampen latent anti-government discourse that might offer early warnings of impending instability.

6.4.7.1 Mitigation

As noted in the previous sections, such self-censorship applies only to the country enforcing it, and when the global media is pooled together, early warning signals are likely to be manifest in the media of other nations.

6.4.8 THE PARADOX OF PREDICTION AND SELF-FULFILLING PROPHECIES

In closing, it is worth noting a critical limitation of the results of this dissertation that could potentially arise if the results were operationalized into a policymaking environment: the paradox of prediction and the self-fulfilling prophecy. If the system were to generate a forecast that suggested a greater risk of instability in a particular country and policymakers were to subsequently take military or diplomatic action based on that forecast to prevent that instability, they would cause the forecast to be wrong. In essence, the purpose of conflict forecasting is so that action can be taken to change the conditions in order to prevent the forecast. This makes it difficult to evaluate the results of the model: if peacekeeping troops were deployed across a country rated at high risk for a coup, and the country subsequently does not undergo a coup, was the original forecast wrong or did the presence of the massive number of foreign troops cause the potential coup participants to rethink their strategy? It also presents problems for self-learning models that incorporate a feedback loop to evaluate changes in the link between latent indicators and subsequent physical action. If action is taken based on the forecast, this becomes a violation to the observational closed-system loop expected by forecasting systems, and will skew its understanding of the predictive power of the various indicators. Finally, the opposite can be true: if a model forecasts a strong likelihood of collapse in a country, policymakers and military planners may withdraw nation-building funds and military support until after the political situation has stabilized, accelerating the country into a collapse that it may not have undergone if that support had remained. In other words, forecasts can cause self-fulfilling prophecies as external actors take action based on the forecasts that may cause them to occur.

6.4.8.1 Mitigation

Until large-scale interventions, such as military action or nation building efforts, are conducted based primarily on the forecasts of early warning systems, such paradoxes are entirely theoretical. Of all of the limitations outlined in this section, this is the only one that is not easily remediable, since by its very nature, it represents a concerted "insider attack" designed to thwart the accuracy of the algorithms' output by taking action to invalidate it. Until such a system were to be deployed in production with its forecasts leading to interventions, it is difficult to assess the impact this might have. Most likely, however, these interventions will result in new media coverage reflecting the changing environment, which would feed back into the system to adjust the forecasts as the impacts of the interventions solidify.

6.5 CONTRIBUTION TO THE LITERATURE

The vast body of previous work in this area over more than four decades, coupled with the enormous sums of funding (over $125M in the last three years alone) expended on past efforts that yielded limited outcomes, stands testament to the impact that answers to the research questions outlined in this dissertation could present. The evaluative framework for latent forecasting created through this work will significantly enhance the ability of scholars to explore the latent-physical linkage. The ability to develop new quantitative latent forecasting models using this framework will also substantially contribute to the national security of the United States, the ability of US policymakers to more efficiently target conflict prevention resources abroad, and offer a powerful new empirical test bed for the social sciences and humanities.

6.5.1 NATIONAL SECURITY AND CONFLICT PREVENTION

Models capable of successfully forecasting impending physical unrest around the world would constitute an enormous contribution to the national security of the United States. Through the advance identification of at-risk locations and emerging global hotspots, US policymakers will be able to more precisely target conflict prevention resources, including nation building efforts, to most effectively mitigate those threats to peace. New theories and descriptive understandings of the media behaviors and driving forces of global-scale social systems will ultimately allow for realtime synthesis of global restlessness and enable monitoring and analytical exploration of those trends. Better understanding of the latent indicators offering the greatest predictive insight and their geographic and cultural variations will offer profoundly new insights into the ways in which group unrest manifests itself in latent signals in the global media and expand current understanding of social systems beyond limited case studies, towards models that incorporate the nonlinear clustering and feedback effects that define human interaction. The ultimate ability to apply these models to calculate such measures in realtime on live data streams will greatly enrich fragility indexes and political risk calculations by incorporating a rich strata of perceptual cues capturing emergent multi-scale behaviors from the micro-level of the individual through the macro-level of the entire planet.

6.5.2 A NEW EMPIRICAL TESTBED FOR THE SOCIAL SCIENCES AND HUMANITIES

This dissertation has presented a novel approach to the study of the socially-constructed world by facilitating the transition of latent-based social conflict research from small-n case studies towards new analytical approaches capable of integrating the media-action link and modeling real-world media and

social environments at a scale from global to local. Understanding of the interplay between the media environment and social-political interactions may be enhanced through access to computational modeling techniques capable of representing the latent-physical connection, case studies on the datasets and variables offering the greatest predictive power, and case examples of the scalability characteristics of the computational models, offering new insight into questions such as how the global media ecosystem can capture early warning indicators of emerging unrest.

A significant and immediate contribution of this work is the set of representational structures, algorithmic approaches, and computational methods capable of tractably encoding mass-scale high-resolution global physical behavior even within the confines of a doctoral dissertation, leading to new methods of thinking of social modeling of qualitative data like media content. In particular, the efficacy of generating measurable and robust narrative traces linking large event datasets with quantitative measures of massive text archives using fully automated pattern-detection methods suggests the broader applicability of such data analysis towards quantitative sense-making in an array of data archives and could invigorate its use in the humanities and social sciences.

The ability to quantitatively model predictive latent cues such as those identified in the previous chapter will enable scholars across disciplines including media and communications, political science, and history to move beyond small-n case studies of selected events towards realtime earth observation and forecasting akin to that being pursued by the physical sciences. This in turn will enable for the first time the ability to experimentally test theories of media proxies of group and network dynamics occurring at a global scale over a period of decades. Lessons learned in the challenges, limitations, computational scalability, and variations in the predictive power of specific datasets and methods presented here will yield considerable insight of rich applicability to a wide array of scholarly disciplines.

Finally, the novel framework and classification approach demonstrated here for forecasting future unrest will offer a rich test bed for future work from the broader academic community to situate a wide array of new latent indicators in their disciplinary humanistic and theoretical foundations.

REFERENCES

18 Days at the Center of Egypt's Revolution. (2011, February 12). *New York Times*. Retrieved from http://www.nytimes.com/interactive/2011/02/12/world/middleeast/0212-egypt-tahrir-18-days-graphic.html

2nd LD German President Horst Koehler resigns. (2010, May 31). *Xinhua*.

7.2-magnitude quake hits Indonesia: seismologists. (2010, May 9). *Agence France Presse.*

Ahmed, M., Spagna, S., Huici, F., & Niccolini, S. (2013). A peek into the future: predicting the evolution of popularity in user generated content. *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 607-616). ACM.

Air France crash probe could take 'year and a half'. (2009, August 31). *Agence France Presse*.

AL follows Tunisia's events with hope for stability: Moussa. (2011, January 15). *Xinhua*.

Alesina, A., & Ferrara, E. (2004). Ethnic diversity and economic performance. *National Bureau of Economic Research Report no. w10313*.

Althaus, S., & Leetaru, K. (2011). Do 'We' Have a Stake in This War? A Worldwide Test of the In-Group Out-Group Hypothesis Using Open-Source Intelligence. *International Studies Association 2011 Annual Meeting*. Montreal, Canada, March 16-19, 2011.

Anderson, C. (2008, June 23). The end of theory: the data deluge makes the scientific method obsolete. *Wired*. Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Antonellis, I., Bouras, C., & Poulopoulos, V. (2006). Personalized news categorization through scalable text classification. Paper presented at *Frontiers of WWW Research and Development-APWeb 2006* (pp. 391-401). Berlin: Springer Berlin Heidelberg.

AP Names Torchia as bureau chief in South Africa. (2013, January 3). *Associated Press*. Retrieved from http://www.ap.org/Content/AP-In-The-News/2013/AP-names-Torchia-as-bureau-chief-in-South-Africa

Arab leaders patch up ties over Gaza: report. (2009, January 20). *Xinhua*.

Atkinson, M., & Van der Goot, E. (2009). Near real time information mining in multilingual news. *Proceedings of the 18th international conference on the world wide web*, 1153-1154.

Atran, S., & Ginges, J. (2012). Religious and Sacred Imperatives in Human Conflict. *Science*, 336, 855-857.

Bagozzi, B. & Schrodt, P. (2012). Detecting Latent Topics in Poltical News Reports using Latent Dirichlet Alocation Models. *Proceedings of the European Political Science Association Meetings*, Berlin, Germany.

Bagozzi, B., & Schrodt, P. (2012). The Dimensionality of Political News Reports. *Proceedings of the European Political Science Association Meetings*, Berlin, Germany.

Bali Democracy Forum opens in Indonesia. (2009, December 10). *Agence France Presse*.

Bandari, R., Asur, S., & Huberman, B. (2012). The Pulse of News in Social Media: Forecasting Popularity. *Arvix*. Retrieved from http://arxiv.org/pdf/1202.0332v1.pdf

Barnett, M. (1993). Confronting the costs of war: military power, state, and society in Egypt and Israel. Princeton University Press.

Bean, H. (2011). No More Secrets: Open Source Information and the Reshaping of US Intelligence: Open Source Information and the Reshaping of US Intelligence. Westport, CT: Praeger.

Bell, A. (1991). The language of news media. Oxford: Blackwell.

Bengston, D., & Xu, Z. (1995). Changing national forest values: A content analysis. *Research paper, NC 323*. St. Paul, Minn.: North Central Forest Experiment Station, Forest Service, U.S. Dept. of Agriculture.

Bertrand, J. (2004). Nationalism and ethnic conflict in Indonesia. University of Cambridge: Cambridge, UK.

Bird flu remains a threat: WHO. (2010, March 24). *Agence France Presse*.

Blei, D., & Lafferty, J. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 17-35.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8.

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition, 30*(7), 1145-1159.

Brandt, P., Freeman, J., & Schrodt, P. (2009). Real Time, Time Series Forecasting of Political Conflict. *International Studies Association 2009 Meeting*. Retrieved from http://www.polisci.umn.edu/~freeman/BFS-ISA20090210.pdf

Brandt, P., Freeman, J., & Schrodt, P. (2011). Racing Horses: Constructing and Evaluating Forecasts in Political Science. *28th Annual Summer Meeting of the Society for Political Methodology*. Retrieved from http://eventdata.psu.edu/papers.dir/RHMethods20110720.pdf

Brazil to appoint ambassador to Honduras. (2011, June 1). *Xinhua*.

Brazilian humorists protest ban on political satire. (2010, August 22). *Agence France Presse*.

Brill, E. (1992). A simple rule-based part of speech tagger. *Proceedings of the Third conference on applied natural language processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, 152-155

Britain and Germany back tougher Iran sanctions. (2010, April 1). *Agence France Press*.

Brown, Merkel meet over economy, Iran. (2010, April 1). *Xinhua*.

Car arson lunacy haunts Berlin for 3 nights in a row. (2011, August 18). *Xinhua*.

Cars torched in Berlin for fourth consecutive night. (2011, August 19). *Agence France Presse*.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning,* Pittsburg, PA, 161-168

Cha, A. (2012, June 6). 'Big Data' from social media, elsewhere online redefines trend-watching. The *Washington Post*. Retrieved from http://www.washingtonpost.com/business/economy/big-data-from-social-media-elsewhere-online-take-trend-watching-to-new-level/2012/06/06/gJQArWWpJV_story.html

Chadefaux, T. (2012). Early Warning Signals for War in the News. *Working Paper*. Retrieved from http://www.thomaschadefaux.com/files/EWS.pdf

Chen, C., Wu, K., Srinivasan, V., & Zhang, X. (2011). Battling the Internet Water Army: Detection of Hidden Paid Posters. *Arxiv*. Retrieved from http://arxiv.org/abs/1111.4297

Chiang, C., & Knight, B. (2008). Media Bias and Influence: Evidence from Newspaper Endorsements. *Working Paper, Brown University*. Retrieved from http://www.econ.brown.edu/fac/Brian_Knight/endorsements4.pdf

Chinese embassy in Egypt ready to help compatriots fleeing from unrest-hit Libya. (2011, February 22). *Xinhua*.

Chinese Premier meets Egyptian parliament speaker. (2007, October 25). *Xinhua*.

Chunara, R., Andrews, J., & Brownstein, J. (2012). Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. *The American Journal of Tropical Medicine and Hygiene, 81*(1), 39-45. Retrieved from http://www.ajtmh.org/content/86/1/39.full

Clifton, J. (2012). Middle East Leads World in Negative Emotions. *Gallup*. Retrieved from http://www.gallup.com/poll/155045/middle-east-leads-world-negative-emotions.aspx

Clinton meets Indonesian leader. (2009, February 19). *Agence France Presse*.

Clionadh, R., Linke, A., Hegre, H., Karlsen, J. (2010). Introducing ACLED: An Armed Conflict Location and Event Dataset. *Journal of Peace Research 47*(5), 1-10.

Coburn, A. (2012). The LINGUA::EN::Tagger module. *CPAN*. Retrieved from http://search.cpan.org/~acoburn/Lingua-EN-Tagger-0.23/Tagger.pm

Cran packages by date of publication. (2013, March 1). Retrieved from http://cran.r-project.org/web/packages/available_packages_by_date.html

Crimson Hexagon. (2011, December 8). Twitter and perceptions of crisis related stress. *United Nations Technical Report*. Retrieved from http://www.unglobalpulse.org/projects/twitter-and-perceptions-crisis-related-stress

D'Orazio, V., Yonamine, J., & Schrodt, P. (2011). Predicting Intra-State Conflict Onset: An Event Data Approach using Euclidean and Levenshtein Distance Measures. *Presented at the 69th Annual Meeting of the Midwest Political Science Association*, Chicago, Il, USA.

DARPA FutureMap Program. (2002). DARPA. Page no longer available, archival copy of August 17, 2002 from the Internet Archive used instead, retrieved from http://web.archive.org/web/20020817001632/http://www.darpa.mil/IAO/FutureMap.htm

Davenport, T., & Beck, J. (2001). The attention economy: Understanding the new currency of business. Harvard Business Press.

Deng, L., & Poole, M. (2010). Affect in Web Interfaces: A Study of the Impacts of Web Page Visual Complexity and Order. *MIS Quarterly. 34*(4), 711-730.

Deutsch, K. (1957). Mass communications and the loss of freedom in national decision-making: a possible research approach to interstate conflicts. *Conflict Resolution 1*(2), 200-211

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2008). Misc functions of the Department of Statistics (e1071). *R Foundation for Statistical Computing*, Vienna, Austria.

Doran, C., Pendley, R., & Antunes, G. (1973). A Test of Cross-National Event Reliability. *International Studies Quarterly 17*, 175-203.

Dugas, A., Hsieh, Y., Levin, S., Pines, J., Mareiniss, D., Mohareb, A., Gaydos, C., Perl, T., Rothman, R. (2012). Google Flu Trends: Correlation With Emergency Department Influenza Rates and Crowding Metrics. *Clinical Infectious Diseases*. Retrieved from http://cid.oxfordjournals.org/content/early/2012/01/02/cid.cir883.abstract

Easterly, W. (2000). Can institutions resolve ethnic conflict? *World Bank Publications Vol 2482*.

Egypt evacuates over 13,600 nationals from Lebanon via Syria. (2006, July 25). *Xinhua*.

Egypt hopes to make Alexandria its first smoke-free city. (2010, June 10). *Agence France Presse*.

Egypt says to respect Tunisian people's choice. (2011, January 15). *Xinhua*.

Egypt, Italy Sign Agreement on Security Cooperation. (June 18, 2004). *Xinhua*.

Egypt, Saudi Arabia to strengthen political consultations. (2007, July 30). *Xinhua*.

Egyptian forces kill two smugglers at Egypt-Gaza border. (2005, November 1). *Xinhua*.

Eisenstein, J., O'Connor, B., Smith, N., Xing, E. (2010). A latent variable model for geographic lexical variation. *EMNLP'10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277-1287.

Electronic Market-Based Decision Support: SB012-012. (2001). DARPA. Retrieved from http://www.acq.osd.mil/osbp/sbir/solicitations/sbir20012/darpa012.pdf

Ellemers, N. (2012). The Group Self. *Science*, 336, 848-852.

Embraer wary of reputation in wake of China crash: experts. (2010, August 26). *Agence France Presse*.

Esteban, J., Mayoral, L., & Ray, D. (2012). Ethnicity and Conflict: Theory and Facts. *Science*, 336, 858-865.

Facebook. (2012). United States Securities and Exchange Commission Form S-1. Retrieved from http://www.sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm

Fahim, K. & Stack, L. (2011, January 2). Fatal bomb hits a church in Egypt. *New York Times*. Retrieved from http://www.nytimes.com/2011/01/02/world/middleeast/02egypt.html

Fahim, K., El-Nagger, M., Stack, L., & Ou, E. (2011, February 9). Emotions of a reluctant hero galvanize protesters. *New York Times*. p. A14.

Fearon, J. (2003). Ethnic and Cultural Diversity by Country. *Journal of Economic Growth*, 8, 195-222.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software, 25*(5), 1-54.

Floating Sheep. (2012, July 4). Church or Beer? Retrieved From http://www.floatingsheep.org/2012/07/church-or-beer-americans-on-twitter.html

Foreign aid pours into quake-hit Indonesia. (2009, October 4). *Agence France Presse*.

Foreign Policy. (2012). The Failed States Index. Retrieved from http://www.foreignpolicy.com/failedstates2012

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3, 1289-1305.

Fox, C. (1989). A stop list for general text. *ACM SIGIR Forum, 24*(1-2), 19-21.

French FM arrives in Cairo to discuss Mideast crisis. (2006, July 21). *Xinhua*.

Galtung, J., & Ruge, M. (1965). The Structure of Foreign News The Presentation of the Congo, Cuba and Cyprus Crises in Four Norwegian Newspapers. *Journal of Peace Research*, *2*(1), 64-90.

Gans, H. (1979). Deciding what's news: A study of CBS evening news, NBC nightly news, Newsweek, and Time. TriQuarterly Books.

Gao, J., Hu, J., Mao, X., & Perc, M. (2012). Culturomics meets random fractal theory: insights into long-range correlations of social and natural phenomena over the past two centuries. *Journal of the Royal Society, 9*(73), 1956-1964.

Gerbner, G., & Marvanyi, G. (1977). The many worlds of the world's press. *Journal of Communication, 27*(1), 52–66.

German president quits over military remarks. (2010, May 31). *Associated Press*.

Germany faces national public workers strike. (2010, February 3). *Xinhua*.

Gerner, D., & Schrodt, P. (1996). The Kansas event data system: a beginner's guide with an application to the study of media fatigue in the Palestinian intifada. Presented at the *American Political Science Association meetings*, San Francisco.

Gerner, D., Schrodt, P., Abu-Jabr, R., & Yilmaz, O. (2002). Conflict and Mediation Event Observations (CAMEO): A new event data framework for the analysis of foreign policy interactions. *International Studies Association*, New Orleans.

Ghosn, F., Palmer, G., & Bremer, S. (2004). The MID3 data set, 1993–2001: Procedures, coding rules, and description. *Conflict Management and Peace Science, 21*(2), 133-154.

Gladstone, R., & Afkhami, A. (2012, January 25). Pattern of intimidation is seen in arrests of Iranian journalists and bloggers. *The New York Times*. p. A10.

Gleditsch, N., Wallensteen, P., Eriksson, M., Sollenberg, M., & Strand, H. (2002). Armed conflict 1946-2001: A new dataset. *Journal of Peace Research, 39*(5), 615-637.

Golder, S., & Macy, M. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science, 333*(6051), 1878-1881.

Golding, P., & Elliott, P. (1979). Making the news. London: Longman.

Greeks to Withdraw $1 Billion a Day Ahead of Vote. (2012, June 13). *Reuters*.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101(Suppl 1)*, 5228-5235.

Gruhl, D., Guha. R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. *13th international conference on the world wide web*. (pp. 491-501). ACM.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1-30.

Gupta, A. (1992). The song of the nonaligned world: Transnational identities and the reinscription of space in late capitalism. *Cultural Anthropology*, *7*(1), 63-79.

Halzack, S. (2012, March 5). Bloomberg's Billionaires Index. *The Washington Post*. Retrieved from http://www.washingtonpost.com/business/economy/bloombergs-billionaires-index-a-daily-measure-of-the-worlds-wealthiest-people/2012/03/05/gIQADpansR_story.html?hpid=z4

Hamas denies involvement in Eilat rocket attack. (2010, August 3). *Xinhua*.

Hegre, H., & Sambanis, N. (2006). Sensitivity Analysis of Empirical Results on Civil War Onset. *Journal of Conflict Resolution, 50*(4), 508-535.

Honduras president shuns foreign summit amid boycott threats. (2010, May 6). *Agence France Presse*.

Hunt, K. (2012, May 28). China tightens grip on social media with new rules. *CNN.com*. Retrieved from http://www.cnn.com/2012/05/28/world/asia/china-weibo-rules/index.html

Hunt, W. (1997). Getting to war: Predicting international conflict with mass media indicators. University of Michigan Press.

Huxtable, P. (1997). Uncertainty and Foreign Policy-Making: Conflict and Cooperation in West Africa. *Ph.D. Dissertation*, University of Kansas.

IARPA. (2010). Aggregative Contingent Estimation Program: IARPA-BAA-10-05. Retrieved from http://www.iarpa.gov/solicitations_ace.html

IARPA. (2011). Open Source Indicators (OSI) Program Broad Agency Announcement: IARPA-BAA-11-11.

ICWSM. (2011). ICWSM 2011 Data Challenge. Retrieved from http://icwsm.org/data/index.php

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of computational and graphical statistics, 5*(3), pp. 299-314.

Indonesia calls for vigilance in terrorism fight. (2010, October 18). *Agence France Presse*.

Indonesia denies failures in tsunami aid effort. (2010, November 1). *Agence France Presse*.

Indonesian Islamists protest Obama visit. (2010, November 7). *Agence France Presse*.

Israeli fire kills two militants in Gaza. (2009, March 31). *Xinhua*.

Issenberg, S. (2012, January 13). How cutting-edge text analytics can help the Obama campaign determine voters' hopes and fears. *Slate*. Retrieved from http://www.slate.com/articles/news_and_politics/victory_lab/2012/01/project_dreamcatcher_how_cutting_edge_text_analytics_can_help_the_obama_campaign_determine_voters_hopes_and_fears_.html

Italy condemns blasts in Egyptian Red Sea resorts. (2004, October 9). *Xinhua*.

Italy deports 80 illegal Egyptian immigrants. (2005, October 5). *Xinhua*.

Jackson, P. (1989). Maps of meaning: An introduction to cultural geography. Routledge.

Jordan, A. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 841-848.

Kantrowitz, M., Mohit, B., & Mittal, V. (2000). Stemming and its effects on TFIDF ranking (poster session). In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 357-359.

Kareil, H., & Rosenvall, L. (1984). Factors influencing international news flow. *Journalism Quarterly, 61*, 509-516.

Kay, K., Naselaris, T., Prenger, R., & Gallant, J. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352-355.

Kerbel, J., & Olcott, A. (2010). Synthesizing with Clients, Not Analyzing for Customers. *Studies in Intelligence, 54*(4), 11-27.

King, G., & Lowe, W. (2003). An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. *International Organization, 57*, 617-642

King, G., Pan, J., Roberts, M. (2012, June 16). How Censorship in China Allows Government Criticism but Silences Collective Expression. *Harvard Working Paper*. Retrieved from http://gking.harvard.edu/publications/how-censorship-china-allows-government-criticism-silences-collective-expression

Kirkpatrick, D. (2011, February 17). Egyptians say military discourages an open economy. *New York Times*. Retrieved from http://www.nytimes.com/2011/02/18/world/middleeast/18military.html

Kretchun, N., & Kim, J. (2012, May). A Quiet Opening: North Koreans in a Changing Media Environment. *InterMedia*.

Kuwait reaffirms support to Palestinian cause. (2009, January 27). *Xinhua*.

LaFree, G., & Dugan, L. (2007). Introducing the global terrorism database. *Terrorism and Political Violence*, *19*(2), 181-204.

Lasswell, H. (1927). Propaganda technique in the world war. New York: Alfred A. Knopf.

Lasswell, H. (1971). Propaganda Technique in World War I. MIT press.

Lee, C. (2010). The roots of astroturfing. *Contexts: The American Sociological Association, 9*(1), 73-75.

Leetaru, K. (2010). The Scope of FBIS and BBC Open-Source Media Coverage, 1979-2008. *Studies in Intelligence, 54*(1), 17-37.

Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday, 16*(9-5). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040

Leetaru, K. (2012). Fulltext Geocoding Versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia. *D-Lib Magazine, 18*(9/10). Retrieved from http://mirror.dlib.org/dlib/september12/leetaru/09leetaru.print.html

Leetaru, K., & Olcott, A. (2012). Gaps and ways to improve how populations and social groups can be monitored via journalistic and social media to detect fragility. *National Security Challenges: Insights from Social, Neurobiological, and Complexity Sciences, Multilayer Assessment Program*, Office of the Secretary of Defense.

Leonhardt, D. (2012, July 7). When the Crowd Isn't Wise. *New York Times*. p. SR4. Retrieved from http://www.nytimes.com/2012/07/08/sunday-review/when-the-crowd-isnt-wise.html

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 497-506.

LexisNexis. (online). LexisNexis Academic. Retrieved from http://www.lexisnexis.com/en-us/products/lexisnexis-academic.page

Lott, J. (2010). More guns, less crime: Understanding crime and gun control laws. University of Chicago Press.

Lovins, J. (1968). Development of a stemming algorithm. MIT Information Processing Group, Electronic Systems Laboratory.

Lowe, G. (1966). The Camelot Affair. *Bulletin of Atomic Scientists, 22*(5), 44-48.

Lu, X., Bengtsson, L., & Holme, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences, 109*(29), 11576-11581.

Lyotard, J. (1984). Postmodern condition: A report on knowledge (Vol. 10). University of Minnesota Press.

Major news items in leading Egyptian newspapers. (2004, April 24). *Xinhua*.

Major news items in leading Egyptian newspapers. (2005, October 23). *Xinhua*.

Maloof, M. (2003). Learning when data sets are imbalanced and when costs are unequal and unknown. *ICML-2003 workshop on learning from imbalanced data sets II.*

Markoff, J. (2011, October 10). Government Aims to Build a 'Data Eye in the Sky'. *The New York Times*. p. D1

Marshall, M. (2013). Polity IV Project. Retrieved from http://www.systemicpeace.org/polity/polity4.htm

Martindale, C. (1975). Romantic progression: The psychology of literary history. Washington, D.C.: Hemisphere.

Marwell, G., & Oliver, P. (1993). The critical mass in collective action. Cambridge University Press.

Matt Moore named AP Berlin bureau chief. (2008, March 6). *Associated Press*.

McCarthy, J., McPhail, C., Smith, J. (1996). Images of Protest: Dimensions of Selection Bias in Media Coverage of Washington Demonstrations, 1982 and 1991. *American Sociological Review, 61*(3), 478-499.

Meier, P. (2012). Big Data for Development: Challenges and Opportunities. *iRevolution*. Retrieved from http://irevolution.net/2012/06/05/big-data-for-development/

Melucci, A. (1996). Challenging codes: Collective action in the information age. Cambridge University Press.

Mercado, S. (2001). FBIS against the axis. *Studies in Intelligence*, *11*(Fall-Winter). Retrieved from https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/fall_winter_2001/article04.html

Merkel shocked by Berlin car torching wave. (2011, August 18). *Agence France Presse*.

Merkel, Brown back new sanctions against Iran. (2010, April 1). *Associated Press*.

Michel, J., Shen, Y., Aiden, A., Veres, A., Gray, M., Pickett, J., ... & Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176-182.

Mishne, G., & Glance, N. (2006). Predicting movie sales from blogger sentiment. *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*.

Moeller, S. (1999). Compassion fatigue: How the media sell disease, famine, war and death. Psychology Press.

Montgomery, J. , Hollenbach, F., & Ward, M. (2012). Improving Predictions Using Ensemble Bayesian Model Averaging. *Political Analysis, 20*(3), 271-291.

Mubarak Calls for Redoubling Efforts to Promote Trade Among D-8. (2001, February 25). *Xinhua*.

Mubarak to visit Italy Monday. (2004, October 7). *Xinhua*.

New Brazilian leader to oversee military boost. (2010, October 26). *Agence France Presse*.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641-1646.

O'Brien, S. (2010). Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review, 12*(1), 87–104.

Olcott, A. (2009). Revisiting The Legacy: Sherman Kent, Willmoore Kendall, and George Pettee – Strategic Intelligence in the Digital Age. *Studies in Intelligence, 53*(2), 21-32.

Olcott, A. (2010). Strategies – Good and Bad – for Navigating Information Hyperabundance. *Georgetown University Institute for the Study of Diplomacy Working Papers in New Diplomacy*. Retrieved from http://isd.georgetown.edu/files/Olcott_StrategiesGoodandBad.pdf

Olcott, A. (2012). Open Source Intelligence in a Networked World. Continuum International Publishing: London.

Oliver, P. (1993). Formal Models of Collective Action. *Annual Review of Sociology*, 19, 271-300.

OpenNet Profiles. (2012). Retrieved from http://opennet.net/country-profiles

Ostermeier, E. (2012, January 25). My Message Is Simple. *Smart Politics*. Retrieved from http://blog.lib.umn.edu/cspg/smartpolitics/2012/01/my_message_is_simple_obamas_so.php

Palestinian official says truce with Israel in danger. (2005, June 10). *Xinhua*.

Peters, J. (2010). In a world of online news, burnout starts younger. *The New York Times*. Retrieved from http://www.nytimes.com/2010/07/19/business/media/19press.html

Political Instability Task Force. (2011). Retrieved from http://globalpolicy.gmu.edu/pitf/

Porter, M. (2001). Snowball: A language for stemming algorithms. Retrieved from http://snowball.tartarus.org/texts/introduction.html

Protests continue for second day in Egypt. (2011, January 27). *Xinhua*.

Provalis Research. (2003). WordStat Roget Dictionary. Retrieved from http://www.provalisresearch.com/wordstat/Roget.html

Provalis Research. (2005). WordStat WordNet 2.0 Dictionary. Retrieved from http://www.provalisresearch.com/wordstat/WordNet.html

Radinsky, K., & Horvitz, E. (2013). Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 255-264. Retrieved from http://research.microsoft.com/en-us/um/people/horvitz/future_news_wsdm.pdf

Ramzy, A. (2011, February 21). State Stamps Out Small 'Jasmine' Protests in China. *Time*. Retrieved from http://www.time.com/time/world/article/0,8599,2052860,00.html

Reeves, A., Shellman, S., & Stewart, B. (2006, March). Fair & Balanced or Fit to Print? The Effects of Media Sources on Statistical Inferences. *International Studies Association Conference*. Retrieved from http://web.ku.edu/~keds/papers.dir/Reeves.Shellman.Stewart.pdf

Ricchiardi, S. (2008). Covering the World. *American Journalism Review*. Retrieved from http://www.ajr.org/article.asp?id=4429

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence, 3*(22), 41-46.

Rodriguez, M., Hidalgo, J., & Agudo, B. (2000). Using WordNet to complement training information in text categorization. *Proceedings of 2nd International Conference on Recent Advances in Natural Language Processing II: Selected Papers from RANLP*, 97, 353-364.

Roop, J. (1969). Foreign broadcast information service. *CIA*. Retrieved from https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/foreign-broadcast-information-service/

Rosenblum, M. (1981). Coups and Earthquakes: Reporting the World to America. New York: Harper Colophon.

Roundup: Italy, Egypt boost strategic partnership with 17 new agreements. (2010, May 19). *Xinhua*.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Sarkees, M. & Schafer, P. (2000). The correlates of war data on war: An update to 1997. *Conflict Management and Peace Science, 18*(1), 123-144.

SAS. (2011). Using Social Media and Online Conversations to Add Depth to Unemployment Statistics. *United Nations Technical Report*. Retrieved from http://www.unglobalpulse.org/projects/can-social-media-mining-add-depth-unemployment-statistics

Sayah, R. (2012, May 20). Pakistan blocks Twitter over material deemed blasphemous. *CNN.com*. Retrieved from http://www.cnn.com/2012/05/20/world/asia/pakistan-twitter/index.html

Schelling, T. (1978). Micromotives and Macrobehavior. New York: Norton.

Schrodt, P. & Yonamine, J. (2012). Automated Coding of Very Large Scale Political Event Data. Workshop on New Directions in Text as Data, Harvard. Retrieved from http://eventdata.psu.edu/papers.dir/Schrodt_Yonamine_NewDirectionsInText.pdf

Schrodt, P. (2000). Forecasting Conflict in the Balkans using Hidden Markov Models. *American Political Science Association Annual Meeting*. Washington DC.

Schrodt, P. (2001). Automated coding of international event data using sparse parsing techniques. *International Studies Association Meeting*, Chicago.

Schrodt, P. (2010). Automated production of high-volume, near-real-time political event data. *2010 American Political Science Association Conference*, Washington DC.

Schrodt, P. (2011). Forecasting Political Conflict in Asia Using Latent Dirichlet Allocation Models. *Annual Meeting of the European Political Science Association*, Dublin.

Schrodt, P. (2012). CAMEO Conflict and Mediation Event Observations Event and Actor Codebook V. 1.1b3. Retrieved from http://eventdata.psu.edu/cameo.dir/CAMEO.Manual.1.1b3.pdf

Schrodt, P., Simpson, E., & Gerner, D. (2001). Monitoring conflict using automated coding of newswire reports: a comparison of five geographical regions. In *Conference 'Identifying Wars: Systematic Conflict Research and it's Utility in Conflict Resolution and Prevention', Uppsala* (pp. 8-9). Available online at http://eventdata.psu.edu/papers.dir/KEDS.Uppsala.pdf

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1-47.

Shachtman, N. (2011). Pentagon's Prediction Software Didn't Spot Egypt Unrest. *Wired*. Retrieved from http://www.wired.com/dangerroom/2011/02/pentagon-predict-egypt-unrest/

Shook, E., Leetaru, K., Cao, G., Padmanabhan, A., & Wang, S. (2012). Happy or not: Generating topic-based emotional heatmaps for Culturomics using CyberGIS. *IEEE 8th International Conference on eScience*.

Simon, H. (1971). Designing organizations for an information-rich world. In Greenberger, M., Computers, communication, and the public interest. Baltimore, MD: The Johns Hopkins Press.

Slow return to school for quake-hit Haiti's students. (2010, October 6). *Agence France Presse*.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. *KDD workshop on text mining*, 400, 525-526.

Stierholz, K. (2008). What old news tells us that data does not: The uses of news reports in monetary policy research. *On The Record: A Forum on Electronic Media and the Preservation of News*, New York Public Library, New York City. Retrieved from http://www.crl.edu/sites/default/files/attachments/events/Stierholz-What%20Old%20News%20Tells%20Us%20That%20Data%20Does.pdf

Stolberg, S. (2011, February 16). Shy US intellectual created playbook used in a revolution. *The New York Times*. Retrieved from http://www.nytimes.com/2011/02/17/world/middleeast/17sharp.html

Stoll, R. & Subramanian, D. (2006). Hubs, authorities and networks: predicting conflict using events data. *International Studies Association Annual Meeting*, San Diego, CA.

Stone, P., Dunphy, D., Smith, M. (1966). The general inquirer: A computer approach to content analysis. Cambridge, Mass.: MIT Press.

Sudan FM says summit promotes CPA, resolves outstanding issues. (2010, December 21). *Xinhua*.

Sudan's Beshir to visit Egypt. (2009, March 25). *Agence France Presse*.

Sudanese government and Darfur rebels to resume talks. (2004, October 20). *Xinhua*.

Tankard, J. (2001). The empirical approach to the study of media framing. In Reese, S. D., Gandy Jr, O. H., & Grant, A. E. (Eds.). Framing public life: Perspectives on media and our understanding of the social world, Lawrence Erlbaum, pp. 95-106.

Tatar, A., Antoniadis, P., de Amorim, M., & Fdida, S. (2012). Ranking news articles based on popularity prediction. In *Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 106-110). IEEE.

The science of civil war. (2012, April 21). *The Economist*. Retrieved from http://www.economist.com/node/21553006

Thompson, D., & Borglum, R. (1973). A Case Study of Employee Attitudes and Labor Unrest. *Industrial and Labor Relations, 27*(1), 74-83

Tierney, J. (2010, February 8). Will you be emailing this column? *The New York Times*. Retrieved from http://www.nytimes.com/2010/02/09/science/09tier.html

Todd, J. & Ruane, J. (2009). Ethnicity and Religion. Routledge Handbook of Ethnic Conflict. K. Cordell, S. Wolff, eds., Routledge.

Toffler, A. (1984). Future shock. Bantam.

Top African performers named in latest economic report. (2003, July 31). *Xinhua*.

Troianovski, A. (2010, June 30). China Agency Nears Times Square. *The Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB10001424052748704334604575339281420753918.html

Tunisians rally in support of Egyptian protesters. (2011, January 29). *Xinhua*.

Turkey, Egypt reach consensus on settlement of Israeli-Palestinian conflict. (2004, February 11). *Xinhua*.

Twenge, J., Campbell, K., Gentile, B. (2012). Increases in Individualistic Words and Phrases in American Books 1960-2008. *PLos One. 7*(7). Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0040181

Twin resignations batter crisis-weary Merkel. (2010, June 1). *Agence France Presse*.

Twitter to delete posts if countries request it. (2012, January 28). *CNN.com*. Retrieved from http://www.cnn.com/2012/01/27/tech/twitter-deleting-posts/index.html

US to Chavez: 'Do not repress' your people. (2010, February 4). *Agence France Presse*.

USAID. (2005). Measuring Fragility. Retrieved from http://pdf.usaid.gov/pdf_docs/PNADD462.pdf

Rijsbergen, C. (1979). Information Retrieval. Butterworths, London.

Vandals in Berlin torch 18 cars. (2011, August 17). *Associated Press*.

Ward, M., Greenhill, B., & Bakke, K. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research, 47*(4), 363-375.

Wasley, P. (2012, March 8). Flap jacks, pancakes, or batterbread? *NEH*. Retrieved from http://www.neh.gov/news/flap-jacks-pancakes-or-batterbread

Weigle, B. (2007). Prediction Markets Another Tool in the Intelligence Kitbag: Master of Strategic Studies Degree Strategy Research Project. *US Army War College*. Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA469632

Weng, C., & Poon, J. (2008). A new evaluation measure for imbalanced datasets. *Seventh Australasian Data Mining Conference (AusDM 2008),* 87, 27-32.

Whissell, C., Fournier, M., Pelland, R., Weir, D., & Makarec, K. (1986). A dictionary of affect in language: IV. Reliability, validity, and applications. *Perceptual and Motor Skills, 62*(3), 875-888.

Wigle, J. (2010). Introducing the worldwide incidents tracking system (WITS). *Perspectives on Terrorism*, *4*(1).

Williams, L. (2012). Prediction Markets: Theory and Applications. Routledge.

Wilson, A. (2006). Development and application of a content analysis dictionary for body boundary research. *Literary and Linguistic Computing, 21*(1), 105–110.

Wong, F., Sen, S., Chiang, M. (2012). Why Watching Movie Tweets Won't Tell The Whole Story. *2012 ACM workshop on Workshop on online social networks* (pp. 61-66). ACM. Retrieved from http://arxiv.org/abs/1203.4642v1

Woolley, J. (2000). Using Media-Based Data in Studies of Politics. *American Journal of Political Science, 44*(1), 156-173.

Worrell, M. (2011). Why Nations Go To War. Routledge: New York, NY.

Wu, H. (2006). Systemic determinants of international news coverage: A comparison of 38 countries. *Journal of Communication, 50*(2), 110-130.

Yang, J., & Leskovec, J. (2010). Modeling Information Diffusion in Implicit Networks. *IEEE International Conference On Data Mining (ICDM).* Retrieved from http://cs.stanford.edu/people/jure/pubs/lim-icdm10.pdf

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *22nd annual international ACM SIGIR conference on Research and development in information retrieval*.

Yang, Y., & Pedersen, J. (1997). A comparative study on feature selection in text categorization. *Machine Learning International Workshop*.

Yin, C. (1996). Equilibria of collective action in different distributions of protest thresholds. *Public Choice, 97*, 535-567.

Zelaya supporters detained in Honduras crackdown. (2009, September 30). *Agence France Presse*.