CAN WE FORECAST CONFLICT?

A Framework for Forecasting Global Human Societal Behavior Using Latent Narrative Indicators

Kalev Leetaru – kalev.leetaru5@gmail.com

Vision

- Use historic patterns and current observational environment to forecast future societal unrest.
- Would transform the ability to mitigate violent unrest, enhance first responder impact, and enable new data-driven study of the transformation of grievance to unrest.
- Examine present latent environment to forecast future physical unrest (transition of latent unhappiness to physical action).

Differential Access to Information

- Repressive governments have long forecasted conflict by tapping phones, interrogating citizens, surveillance and using other means to access information.
- This levels playing field for protesters, democracy advocates, and others who live in a relative information vacuum. Platforms like Ushahidi and crowdmapping have been created to centralize open source (news) access.
- Human analysis only available to governments, automated can be done by anyone using Amazon cloud and Google News.
- Weather forecasting: only governments, now anyone.
- Essentially "cultural weather forecasting"
- We'll revisit at the end in discussion

Genesis

- Started off as a dissertation about forecasting, moving from whole-country collapse to day-to-day unrest
- Discovered huge disconnect between physical event and latent forecasting literatures and lack of same-corpus latent+event data
- Worse, most day-to-day unrest databases cover few locations and time periods and are massive initiatives ($38M for one project alone)
- Could a doctoral dissertation actually study day-to-day conflict?

A Framework

- Thus, this dissertation shifted to focus on a framework for evaluating latent forecasting models
- Prove out feasibility of acquiring and converting large archives of news content into physical event records and using them to test latent forecasting models – can it be done as part of a dissertation, not a $38M initiative?
- Prove out framework by testing on "forecasting as classification"

PREVIOUS WORK

THE SCIENCE OF CONFLICT FORECASTING

The Science of Conflict Forecasting

- 40 years of formal “conflict early warning” work
- Poor outcomes: 90% failure rate
- >$250M unclassified over just past decade
- Three primary approaches
  - Human Assessment
  - Prediction Markets
  - Extrapolating

Solitary Expert: Human Assessment

- Lone expert assesses all available information and renders a judgment.
- Most common and oldest, used by nearly all current fragility indexes.
- Wrong more often than right. Personal biases, flawed interpretation, and influence of policymakers can all impact results.

Pooling Experts: Prediction Markets

- Pool multiple experts together to average their assessments: combine their expertise and mitigate impact of individual biases. Can be informal task force like PITF or formal prediction market like FutureMAP.
- Requires active participation, meaning number of viewpoints is very low. Early warnings often manifest outside participant's area.
- Politically sensitive: FutureMAP aborted because of public perception of private individuals profiting from violence.

Data-Driven Physical Extrapolation

- Use data-driven computer models to forecast future unrest based on patterns of past unrest. Highly accurate with prediction of consumer preferences.
- In political realm, use "event databases" like ICEWS.
- By the time physical violence is surging sufficiently for model to detect, its already "too late". Models fail to forecast 90-100% of civil wars and have 70% false positive rate.
- $125M DOD system forecasted just 25% of whole-country collapses and missed Arab Spring entirely.

MOVING FORWARD

FORECASTING USING LATENT MEDIA INDICATORS

Forecasting Using Latent Media Indicators

- Long history of using the news media (and now social media) as a proxy for the "pulse" of society and an indicator for latent unrest (OSINT).
- Remote population assessment – Consumption is truest indicator, but we can only measure Production easily so use it as a proxy.
- Perception matters more than fact. Often too hard to measure the infinite number of driving factors of unrest and many are unobservable. Instead, measure their combined "impact" through change in emotional response of population.

Measuring Emotion

- Economics literature showing that greatest predictive power comes from examining latent dimensions of text: composite emotional response evoked from entirety of current information environment plus personality.
- Use automated "sentiment mining" or narrative analysis to measure emotional content of text.
- Long list of validation studies showing this closely matches human emotional state.

The Big Data Revolution

- 2.5 quintillion new bytes of data each day
- 400M tweets, 6.1T txt msgs, 2 billion new Facebook actions every single day
- Digital revolution means majority of human conversation is mediated or captured through digital channels and can therefore be processed by computers.
- Enables automated data mining at massive scale.
- Numerous studies leveraging this "big data" approach to study human communication and its impact on behavior.

RESEARCH QUESTIONS

OPERATIONALIZING LATENT UNREST FORECASTING

Research Questions

- 1) What are the latent signatures that precede physical societal-scale behavior and manifest themselves in the media in a measurable way?
- 2) Are signatures universal across geographies, or keyed to each location and culture?
- 3) Are signatures universal across classes of physical behavior and intensity levels?

METHODOLOGY

QUANTIFYING RHETORIC AND REALITY

Requirements

- 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.
- Need a cross-national longitudinal quantitative database of physical unrest, together with a collection of latent linguistic indicators measuring the discourse preceding those behaviors.
- Global scale and long time horizon indicates analytic methods must be computationally-driven.

Requirements: Part 2

- 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
- 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)

METHODOLOGY

QUANTIFYING RHETORIC AND REALITY: RHETORIC

The News Media

- The mainstream news media has been a standard for cataloging both human activity and emotional responses since at least World War I and is the basis of Open Source Intelligence (OSINT).
- Long history of study of media and conflict, including latent undercurrents: Karl Deutcsch’s 1957 theory that governments prime their citizens for conflict through the media.
- Used since the 1960’s to catalog human behavior.

Potential News Sources

- Vast array of news collections to pick from:
  - 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)
  - Google News (Chadefaux, 2012)
  - **LexisNexis Academic Universe**

LexisNexis Academic Universe

- Most widely-used news aggregator in academic research.
- More than 5,000 sources stretching back 40 years
- Fulltext of all articles can be downloaded 500 at a time by hand for offline analysis

LexisNexis Academic Universe

- LexisNexis is the defacto standard for the quantitative study of societal-scale political behavior.
- Most common newswires used: Agence France Presse, Associated Press, Xinhua, Reuters, United Press International, and BBC Monitoring (Schrodt, 2010; Reeves, Shellman & Stewart, 2006).
- Contract change means Reuters no longer available.
- Literature suggests patterns may be source-specific, while BBC Monitoring is an aggregator of many sources.

Source Data

- Ultimately chose all international news coverage:
  - Agence France Presse
  - Associated Press
  - Xinhua
- 4,779,821 articles / 1,351,451,029 words
- Even though later analyses focus on 1999-present, had to process all historical data to determine coverage thresholds for different countries
- Limited to just these three sources only to tease out source-specific patterns: top three sources representing each region
- Significant volume of content, received written authorization of Reed Elsevier LexisNexis Group

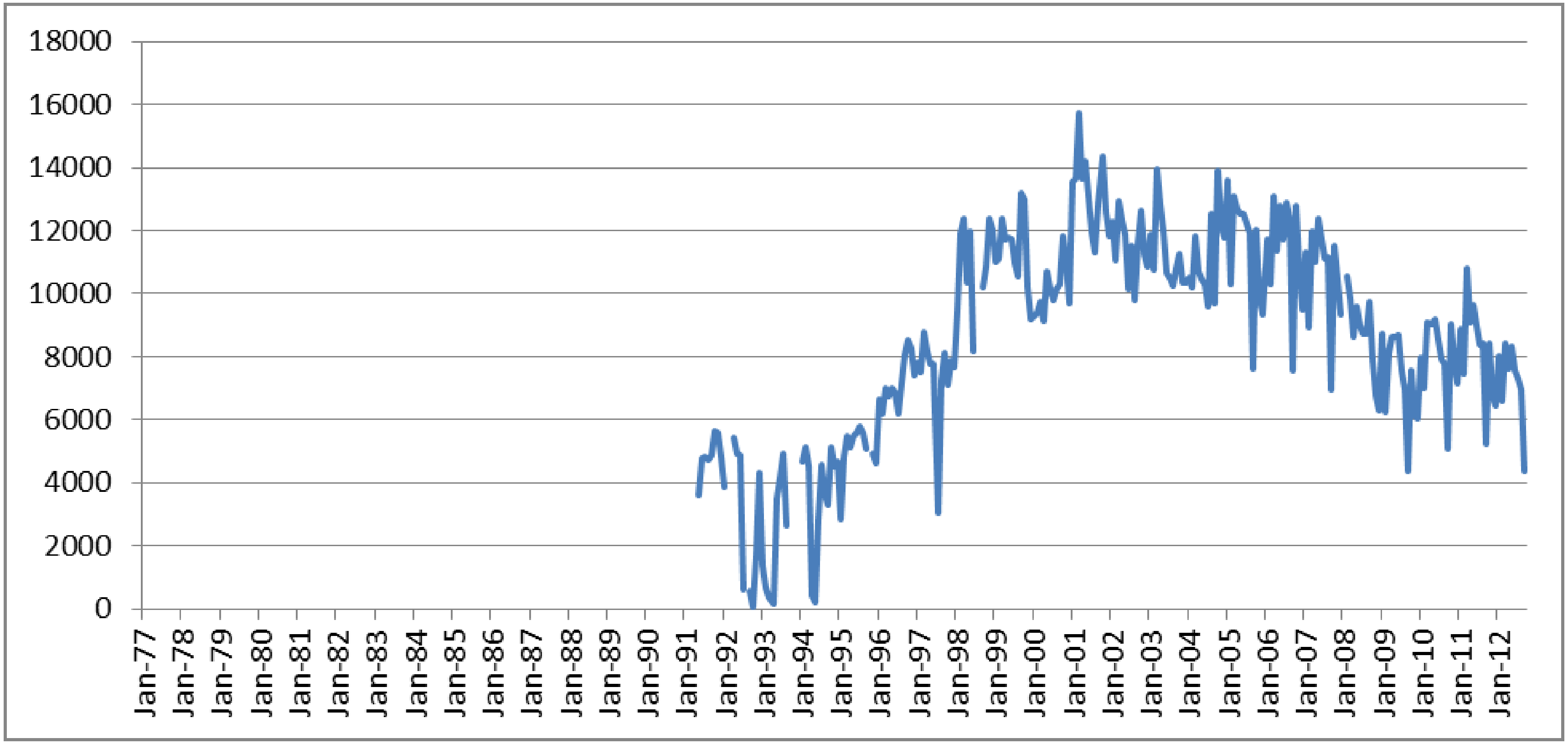
Agence France Presse

- May 1991-Present
- 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).
- “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)

Agence France Presse

- Exclude domestic coverage to reduce bias and sports and economic coverage.
- Final query:
  - *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)*

Agence France Presse



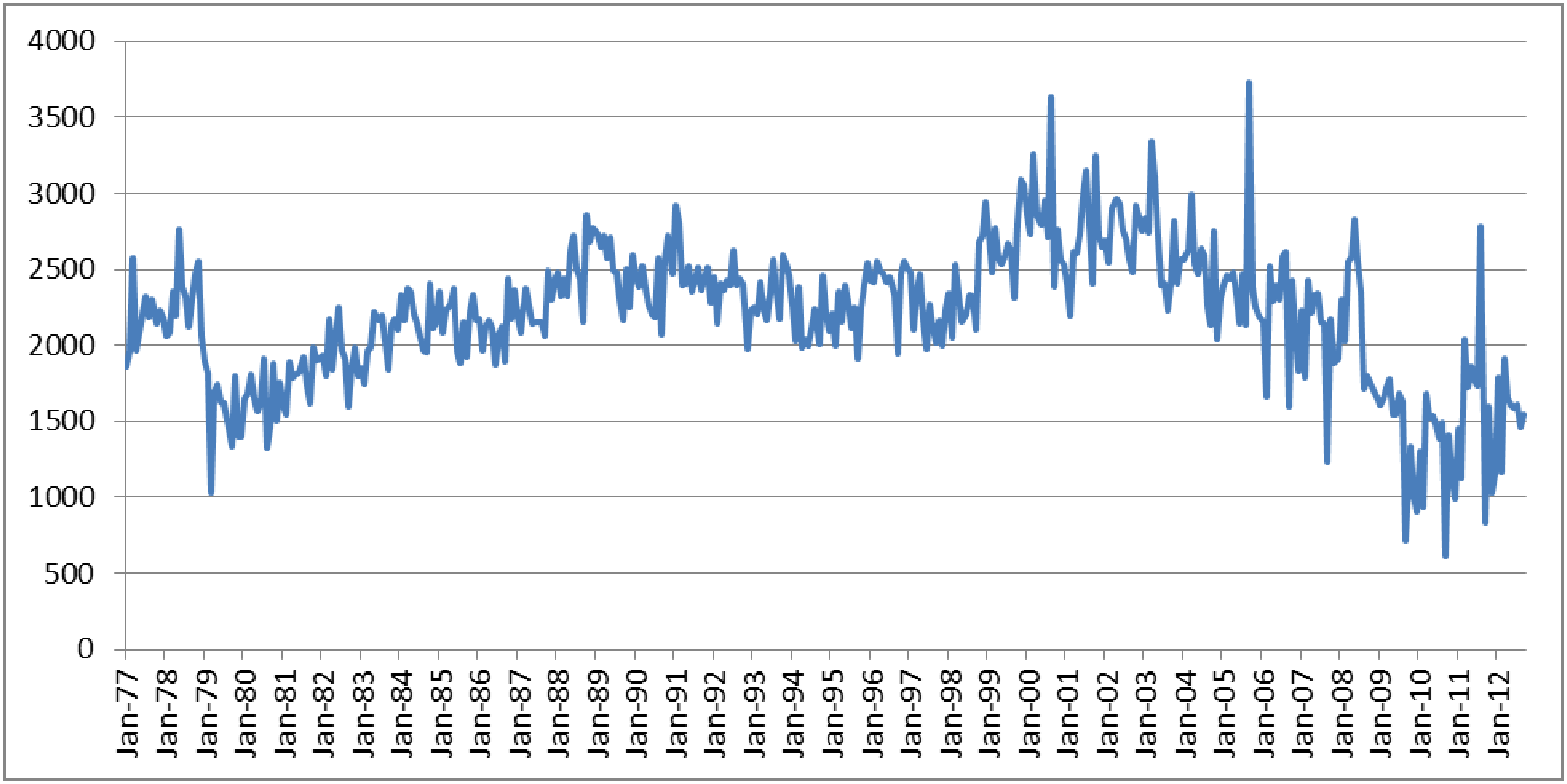
Associated Press

- 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.
- "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." (LexisNexis)

Associated Press

- Primary focus on domestic coverage, so exclude except for major events.
- Introduced SECTION() tag in December 1978, allowing narrowing to just international coverage
- AP Worldstream, which is the main AP international newswire, ends in LexisNexis in 2010
- Query:
  - *"top news" or section(international)*

Associated Press



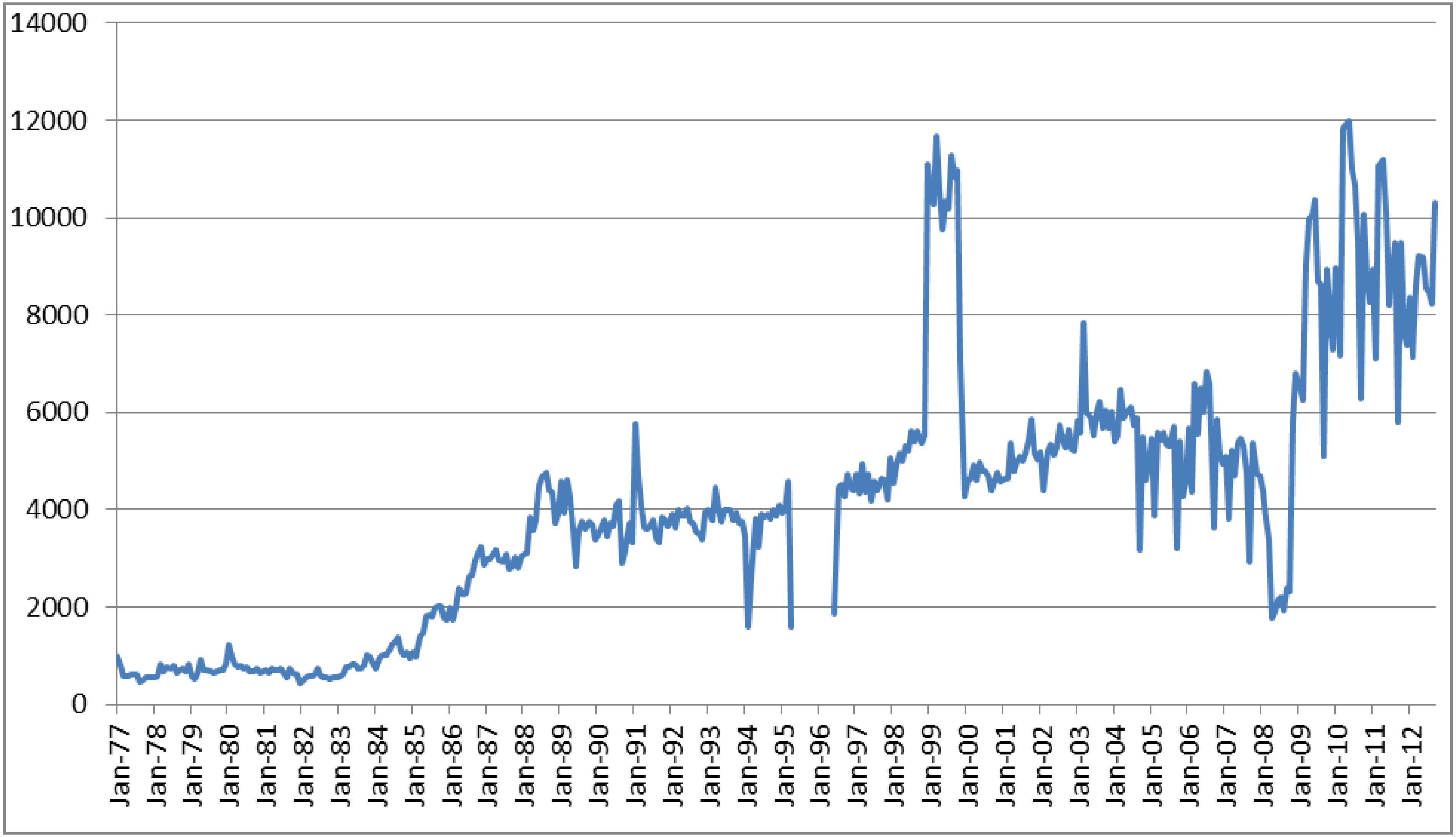
Xinhua

- 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, it has vastly expanded since its founding in the 1931 towards a general-purpose global news service competing with services like Reuters (Troianovski, 2010).
- "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" (LexisNexis)

Xinhua

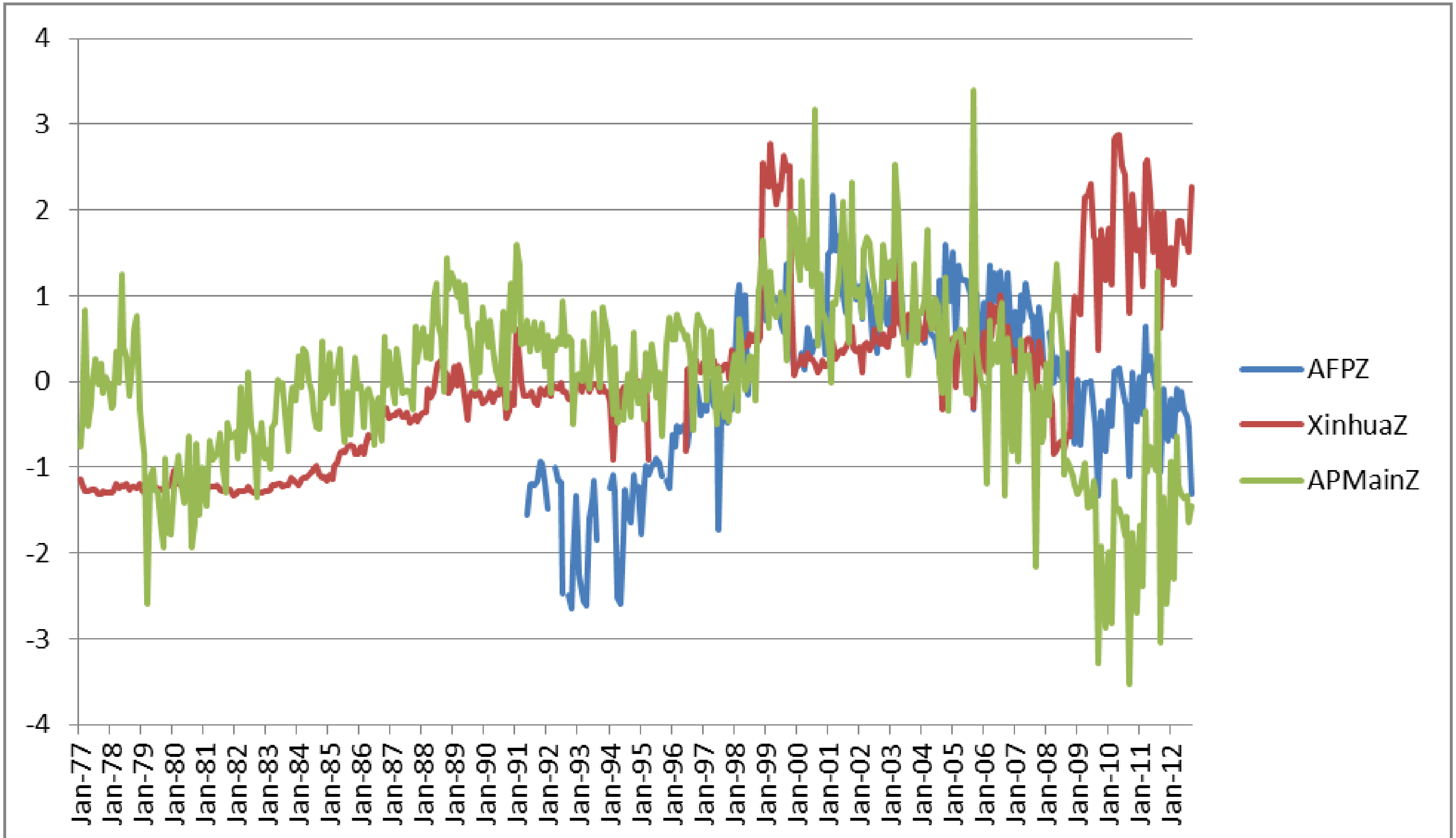
- Primary focus on domestic news, exclude that news. Separate economic newswire, so no need to filter economic news out.
- Massive surge 1998-1999 US Iraq involvement
- Several outages
- Query:
  - *NOT china AND NOT Chinese AND NOT olympic AND NOT snooker AND NOT boxing AND NOT hockey AND NOT marathon 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*

Xinhua

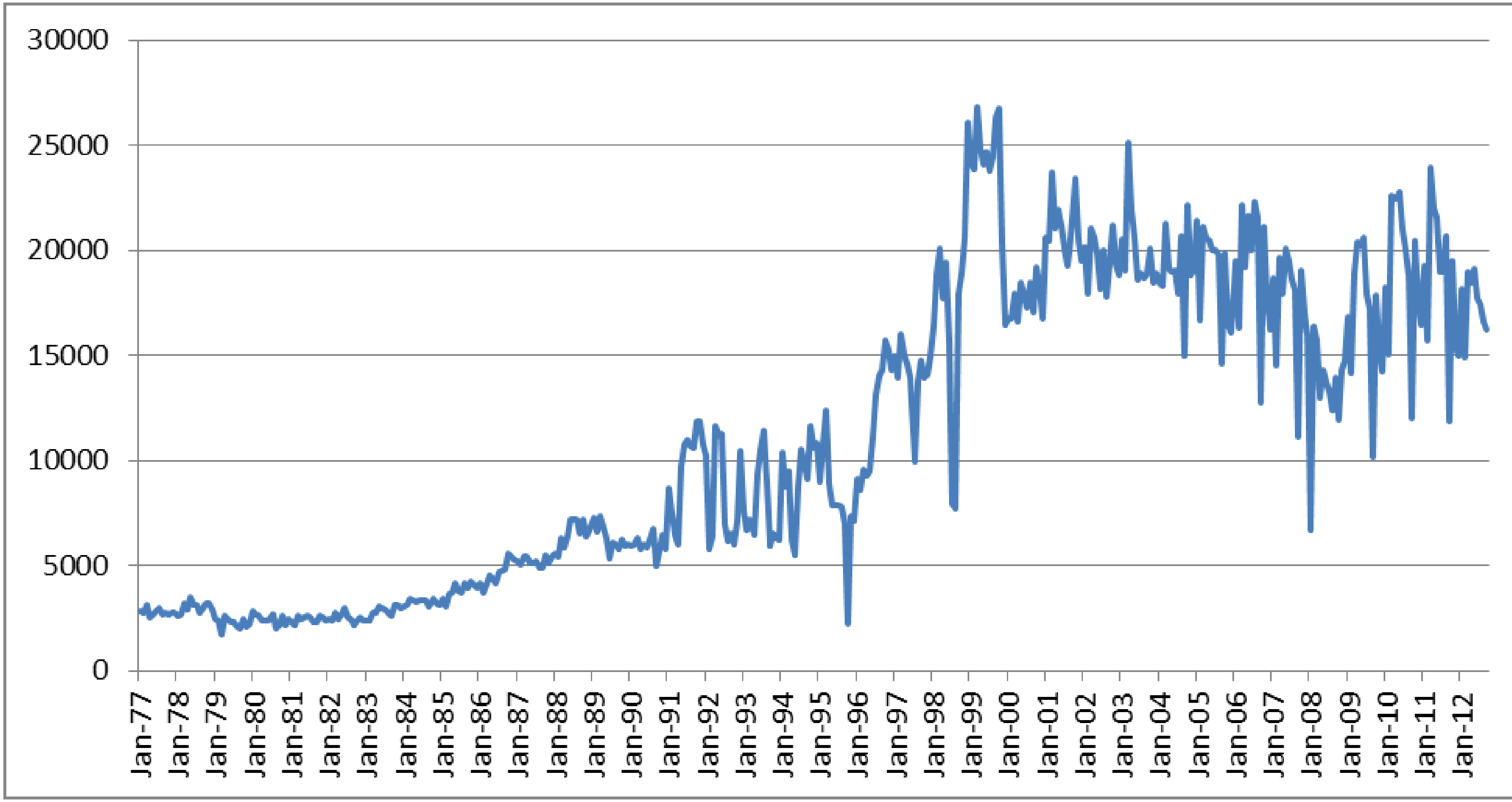


14000
12000
10000
8000
6000
4000
2000
0
Jan-77
Jan-78
Jan-79
Jan-80
Jan-81
Jan-82
Jan-83
Jan-84
Jan-85
Jan-86
Jan-87
Jan-88
Jan-89
Jan-90
Jan-91
Jan-92
Jan-93
Jan-94
Jan-95
Jan-96
Jan-97
Jan-98
Jan-99
Jan-00
Jan-01
Jan-02
Jan-03
Jan-04
Jan-05
Jan-06
Jan-07
Jan-08
Jan-09
Jan-10
Jan-11
Jan-12

Comparing Volume



Combined Volume Profile



Geographic Focus (r=0.95)

Country	% AFP
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

Country	% AP
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
mexico	1.04
south_africa	0.98
poland	0.97
turkey	0.97

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

Post Filtering

- Sports and economic language often similar to conflict language: “battling it out” or “under seige”
- Post processing filter:
  - *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*

METHODOLOGY

QUANTIFYING RHETORIC AND REALITY: REALITY

Codifying Societal Behavior

- How to translate this pile of news text into quantitative discrete measures of human behavior?
- 1960's saw rise of theoretically-informed taxonomies of political behavior such as diplomatic exchanges, protests, and violent conflict
- Allowed humanistic concept of a "protest" to be described in the quantitative terms of its attributes, such as location, date, and actors involved
- Paved way for modern computational modeling

Existing Event Databases

- Event databases very popular, wide array of freely available ones:
  - ACLED (50,000 events for Africa over past decade)
  - GTD (98,000 global terrorist attacks)
  - WITS (USGOV-confirmed terrorist events)
  - CSCW (>25 annual formal battlefield deaths)
- Turnkey systems that combine news + events:
  - European Media Monitor
  - US DOD ICEWS

Existing Event Databases

- Most event databases have few categories, primarily focus on formal battlefield conflict or terrorism
- Limited coverage outside of Africa (or Middle East for terrorism) and short time horizons
- If poor forecasts, can't tell if the limitation is poor algorithm/method or that the event database relied on different sources that covered different events than model input (ie ACLED's use of gov reports for coding vs news-based latent indicators).

ICEWS

- Integrated Conflict Early Warning System, created by DARPA
- Closest to needs of latent forecasting, includes both news and events in over 300 categories
- Only available for US military operational use
- However, the actual event coding core of ICEWS is the open source TABARI system (formerly KEDS) by Philip Schrodt at PSU, available for free download from his website

TABARI and CAMEO

- Most widely-used event coding system today.
- Extensive accuracy benchmarks – only system to pass ICEWS tests
- More than 300 categories of societal behavior
- Codes dyadic events: Actor1 performed Action on Actor2 and date and location (city level) of Action
- Includes self-dyads and non-dyadic events
- Captures all dyadic permutations for peace summits, etc
- Fully autonomous coding system operating in unattended batch mode: input is news and output is events

TABARI and CAMEO

- All events coded into Conflict And Mediation Event Observations (CAMEO) event taxonomy: most widely-used taxonomy today.
- Version 1.1b3 used here 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.”

TABARI and CAMEO

- 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

TABARI and CAMEO

- Many categories have few events.
- Quad Classes:
- 1-5: “Verbal Cooperation”
- 6-9: “Material Cooperation”
- 10-14: “Verbal Conflict”
- 15-20: “Material Conflict”

TABARI and CAMEO

- Combination of TABARI and CAMEO means ANY collection of news articles can be rapidly converted into high-resolution codified event records fully automatically
- Leverages standard political science theoretic taxonomy with over 300 categories to enable very fine-grained study of human behavior
- Events and narrative can come from same text base
- FAST and EFFICIENT: Can run on a laptop and process billions of words in a few hours

TABARI and CAMEO

- All systems, human and machine have error, machines and humans have similar error rates for event coding.
- TABARI the **ONLY** system to pass ICEWS accuracy benchmarks
- Used today as the US DOD operational watchboard that compiles a daily list of all political events worldwide for US military and intelligence analysts

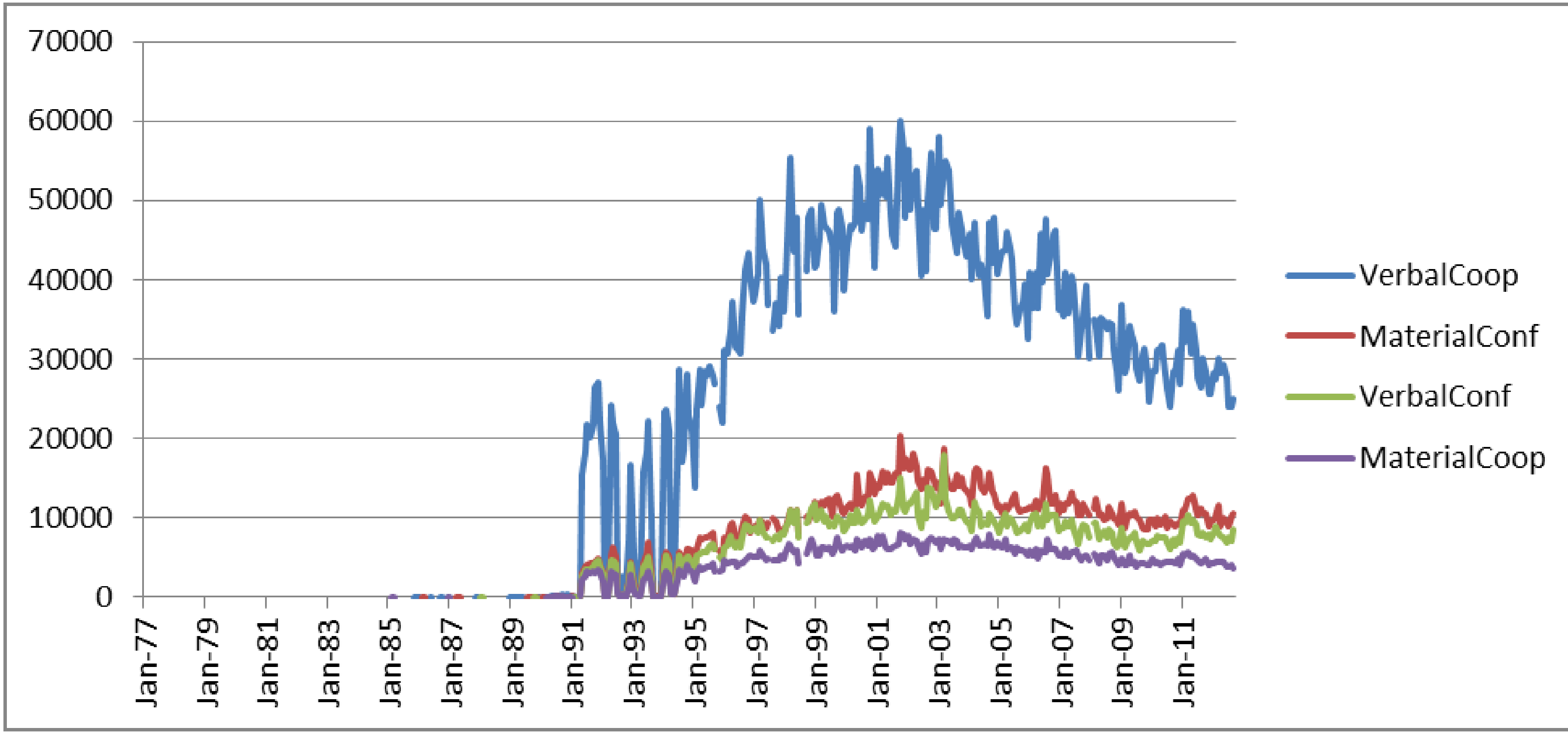
Processing Pipeline

- Content downloaded from LexisNexis by hand, 500 articles at a time via web interface
- Post-processing filter to remove remaining sports and economic news
- TABARI is run on the text to extract all events, operates in full-story mode to extract all events, not just lead sentence events
- Multiple mentions of the same event deduplicated

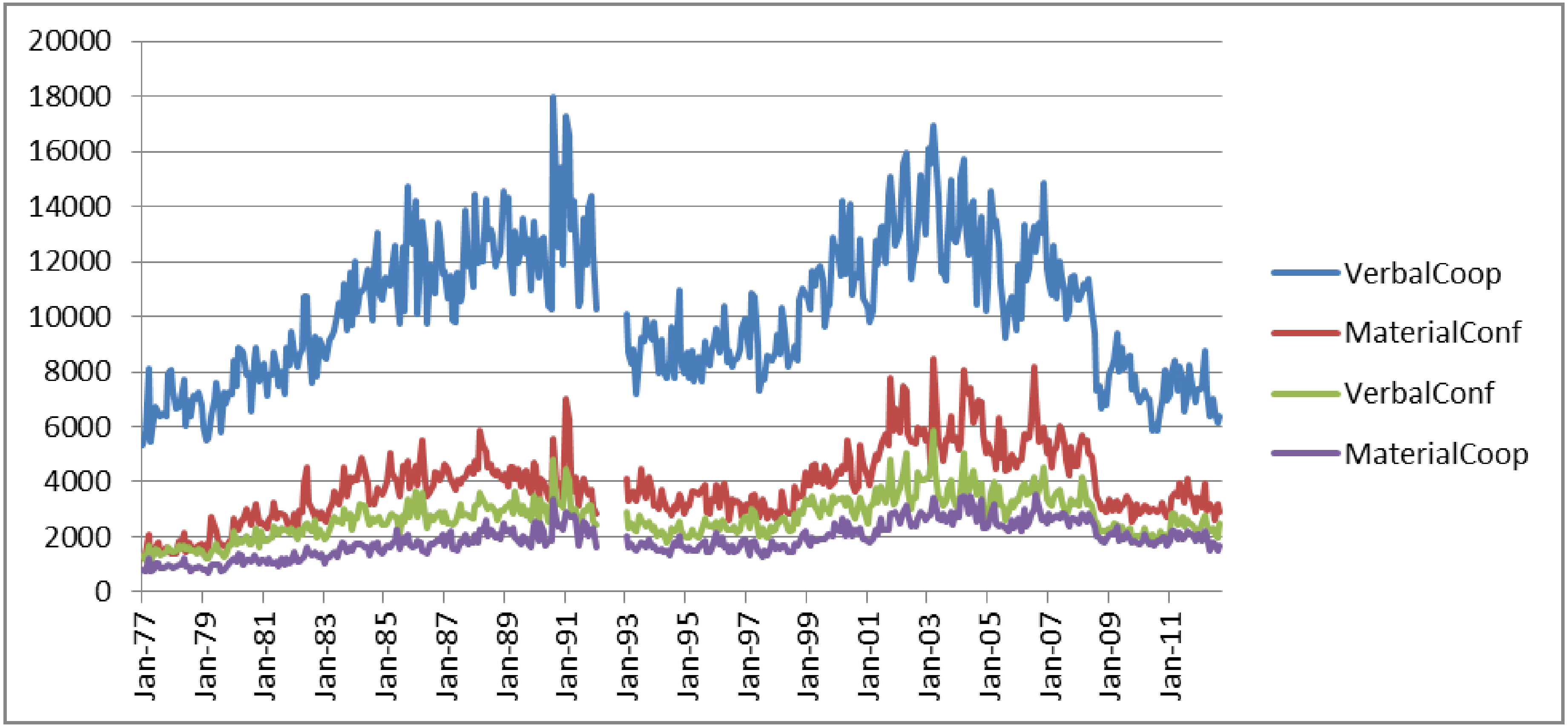
Event Database

- In all 28,877,172 events identified and coded
- Agence France Presse: 14,433,748 (46 words/event)
- Associated Press: 7,811,104 (50 words/event)
- Xinhua: 6,632,320 (50 words/event)

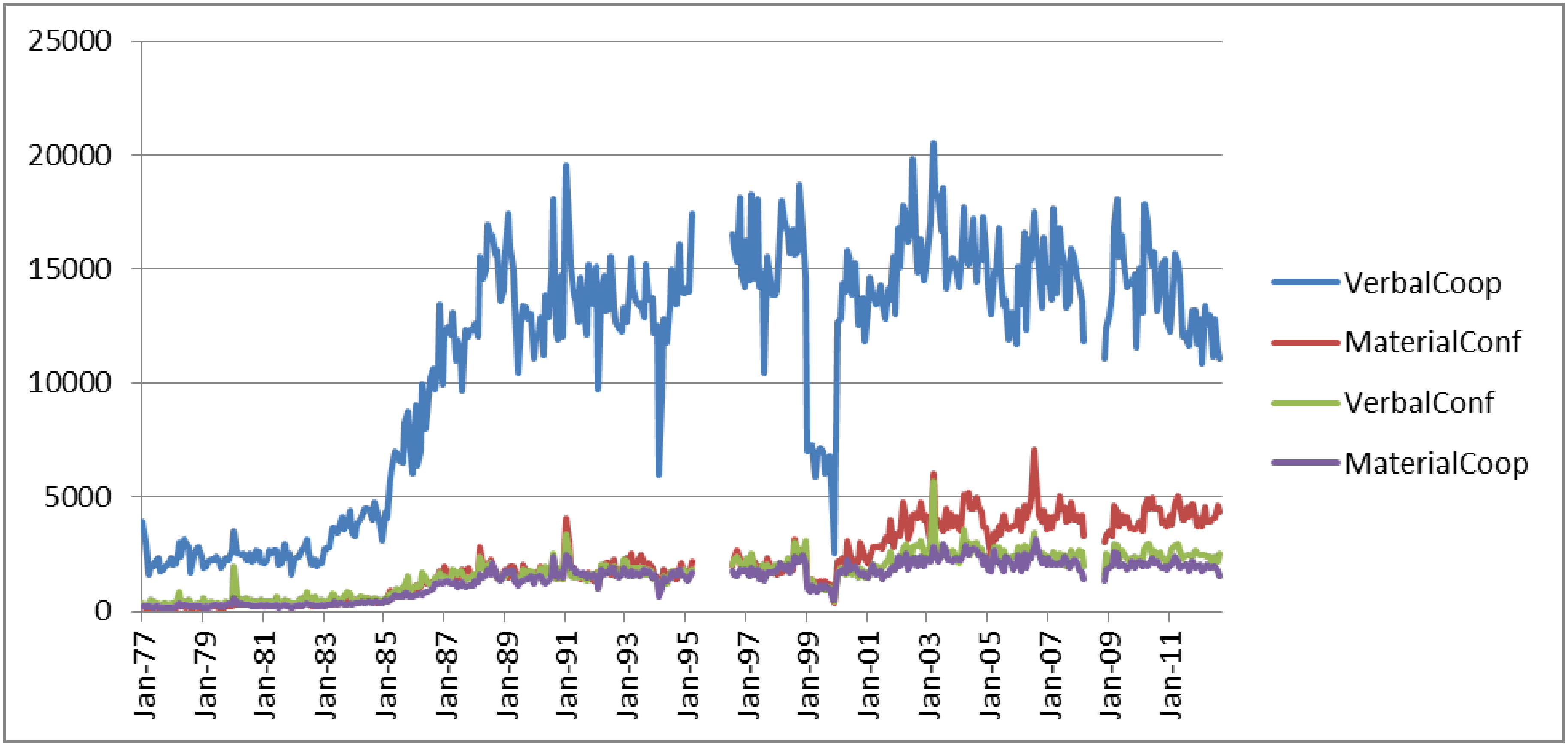
Events: Agence France Presse



Events: Associated Press



Events: Xinhua



Breakdown by Event Type

	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

A FORECASTING FRAMEWORK

APPLYING THE FRAMEWORK TO LATENT FORECASTING

Disconnect

- Political scientists studying physical event patterns construct fine-grained event databases
- Latent forecasters use only gross-scale events and comment on the lack of available fine-grained data (Radinsky & Horvitz do not even know event databases exist)
- Connect the two fields, applying physical event databases to latent forecasting

Current Latent Approaches

- Forced to make use of extremely coarse definitions of unrest and could only focus on physical conflict or other mortality events like disease outbreaks due to the lack of significant cross-national databases of low-intensity events such as peaceful protests or positive actions such as peace accords or aid promises
- Clustering New York Times articles on disease and violent riots into “storylines” (Radinsky & Horvitz, 2013)
- Examining militarized disputes (Chadefaux, 2012)

LexisNexis + TABARI

- Historically TABARI has been used to compile physical event data for physical forecasting, but can it be used to provide a baseline for latent forecasting?
- Download a news archive, run TABARI to construct high-resolution “baseline” of physical behavior, then extract latent dimensions from news and use to forecast that physical behavior

Feasible?

- Even with fully automated coding, largest event dataset available for academic research is 10M records, created by dedicated commercial firm
- ICEWS was $38M project
- Challenges:
  - data acquisition
  - computation needs
  - ease of use
  - adaptability of data

FORECASTING BY CLASSIFICATION

TESTING THE FRAMEWORK

A Vision of Forecasting

- Analyst simply selects a small example list of previous incidents of interest and asks the system to forecast the risk on a day to day basis that similar incidents might occur again at specific time periods in the future
- 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 forecasting future incidences of those events on a day-to-day basis, all without requiring any human assistance beyond selecting the incidents to forecast

A Vision of Forecasting

- Would not require any theoretic understanding of human behavior or expert-based decisions on what data streams and indicators to select
- Supports ad-hoc selection of events of interest
- User simply selects past events of interest and system autonomously learns to forecast future incidents of those events

Text Categorization

- “Text categorization” does exactly this and powers everything from web portals to Google News
- Select small set of example documents of category of interest (such as “basketball” and small set of documents similar but not in the category (such as other sports-related material) as counter-examples.
- System then locates all other “basketball” documents in a larger collection of new documents: train on 100 documents and apply to 100M

Text Categorization

- All documents are converted to "vector space" representation
- Matrix is created where rows are documents and columns are words and cells indicator how many times that word appeared in that document
- Builds probability tables: "player" and "basketball" more often in basketball documents, while "score" and "audience" in basketball and other sports documents, and "football" and "president" very rare in basketball coverage

Text Categorization

- Many algorithms used:
  - Naïve Baysean
  - Support Vector Machines
  - Neural Networks
  - Random Forest
- Naïve Baysean most popular basic classifier due to its relatively high accuracy on a wide array of text, fast execution speed, and simplicity of operation

Forecasting as Classification

- Suggests event forecasting could be treated as a form of text categorization
- Model input is text of news coverage today as of 11:59PM and output is a forecast of the likelihood that different types of events will occur tomorrow and their intensity
- Learns the language most suggestive of future physical behavior
- Generalizes and automates Chadefaux (2012)

Forecasting as Classification

- Since it requires no theoretic understanding of the underlying processes that might cause the events to be forecasted, relying instead purely on learned patterns of lexical probabilities preceding those events (Radinsky & Horvitz, 2013), it can be applied to any type of event, even ones for which little is understood about the underlying driving forces

Forecasting as Classification

- From an information-theoretic perspective, such classifiers are essentially "learning" 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
- Lexical features capture the public information space surrounding 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 (Leetaru & Olcott, 2012).

FORECASTING BY CLASSIFICATION

TESTING THE FRAMEWORK: MODEL CONSTRUCTION

Model Construction

- This dissertation is concerned only with demonstrating feasibility, not constructing the optimal model with the highest forecasting accuracy
- “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).

Discriminative vs Generative

- Discriminative models like Logistic Regression capable of handling term dependence, while generative model of Naïve Baysean assumes term independence (“obama” and “president” vs “onion”)
- Regression would be better-suited to term analysis, but computationally intractable given the size of the total term space and the number of models to be tested (over 1/2M)

Implementation

- R used for all modeling tasks, defacto standard for statistical analysis
- e1071 Naïve Baysean implementation from R, using "SparseM" – not most advanced, but robust and extensively applied to text categorization
- "tm" R package used for term matrixes
- PERL used for text filtering and pre/post-processing
- Use of R allows more than 4,000 other packages to be applied to extend work, while PERL offers future integration to vast array of data ingest services

Tune/Adjust Areas

- Text-Related
  - Number of previous days of text used to forecast the 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

Textual Surrogates

- Five major categories of indicators are used to construct filtered surrogates of the full text: ethnic and religious group-based affiliation, part of speech, emotion, entities, and event-based cues

Ethnic & Religious Conflict

- Ethnic and religious conflict is at the root of a substantial portion of modern societal unrest from Serbia to Sudan (Easterly, 2000)
- 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)

Textual Surrogates: Part 1

- **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.** Given the sometimes-contentious distinction between the boundaries of ethnicity and religion (Todd & Ruane, 2009), 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.

Textual Surrogates: Part 2

- **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.
- **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.

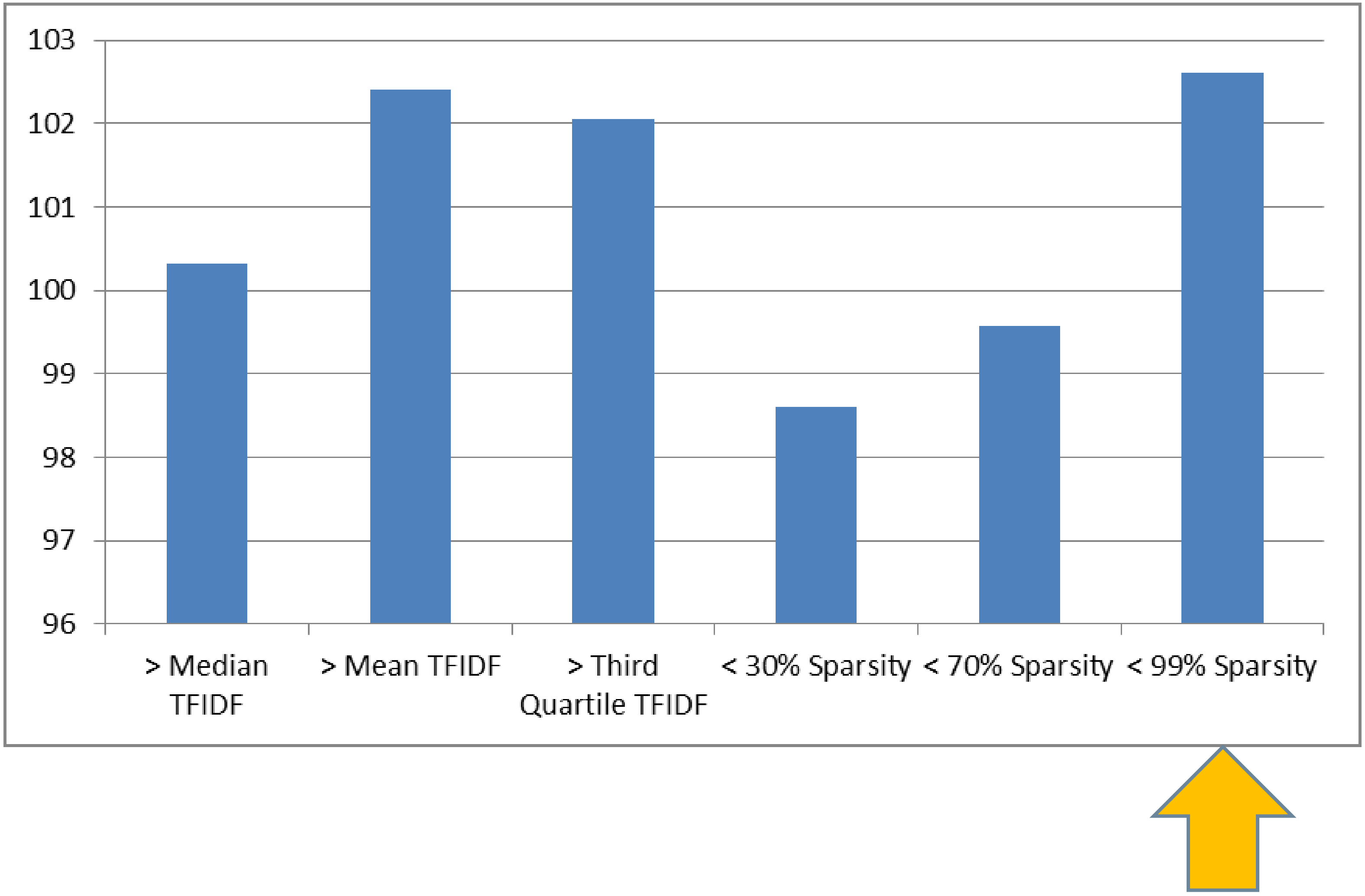
Textual Surrogates: Part 3

- **FTXT_TONE.** This compiles all tonal words, positive or negative (Shook et al, 2012).
- **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. This identifies all locations from cities to local landmarks globally. It includes both the name of the 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. For example, a reference to “French soldiers” will include both the words “French” and “France” in this field.

Textual Surrogates: Part 4

- **META_NAME.** This identifies all person names found in the text. It converts names to single tokens by replacing all spaces and other non-letter characters with underscores, converting "Hosni Mubarak" to "hosni_mubarak".
- **META_ORG.** This identifies all organization names found in the text. It converts names to single tokens by replacing all spaces and other non-letter characters with underscores, converting "United Nations" to "united_nations".

Feature Selection & Weighting



Assessing Accuracy

- Hugely problematic in datasets highly biased towards one category
- Accuracy: correct forecasts – biases towards false positives/negatives (990/1000) = 99% (Weng & Poon, 2008)
- Confusion Matrix: True/False Positive/Negative
- F1 Score, AUC, ROC : Each biases towards high false positive/negative models

Assessing Accuracy

- F1 performs poorly because it aims for a balance between recall and precision for document classification
- Requires decision to penalize false positives or false negatives, but that depends on the situation for forecasting and can't know apriori
- Entire field dedicated to accuracy metrics in imbalanced datasets (Maloof, 2003)
- No single standard metric for event forecasting (Brandt, Freeman & Schrodt, 2011)
- Here, synthetic score of True Positive + True Negative Rate: <100 worse than random, 100=random chance, > 100 better than random, 200=perfect

FORECASTING BY CLASSIFICATION

TESTING THE FRAMEWORK: RESULTS

Xinhua/Egypt

- Start off with Xinhua as the smallest newswire and Egypt, since it has been shown to be highly forecastable
- Start off using single day of text to forecast following day
- Forecast the occurrence of any number of events in any category
- Train on 1999-2005 and forecast 2006-2011

Egypt/Xinhua: >=1 Event

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

Xinhua/Egypt

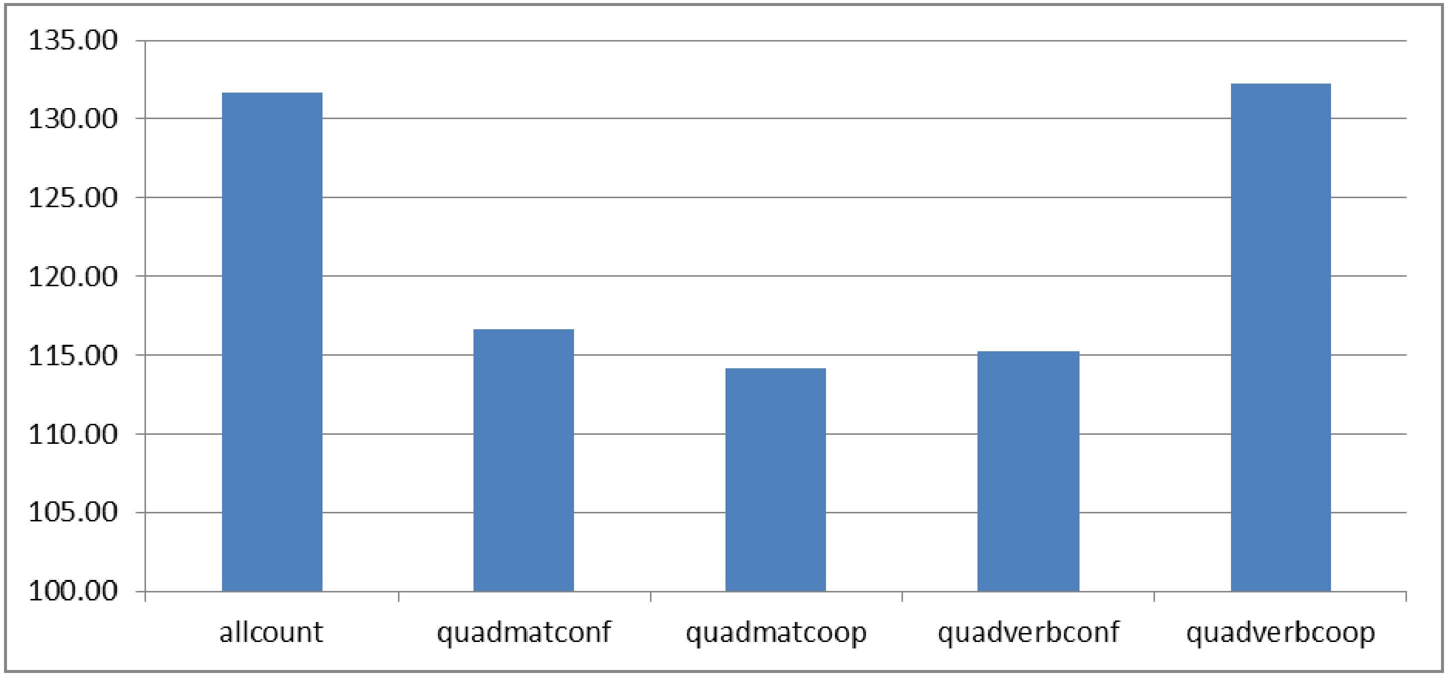
- Instead of forecasting one or more events, forecast “high event” days
- This is what physical event forecasting studies do
- Calculate mean, median, and first and third quartiles of # events per day during training period and use to set threshold for forecasting period
- Intuitively makes sense that it would be easier to forecast massive surge in events rather than isolated event

Xinhua/Egypt: High Event Days

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

Xinhua/Egypt: Event Types

- Are certain classes of events easier to forecast than others?
- Break into Quad Classes

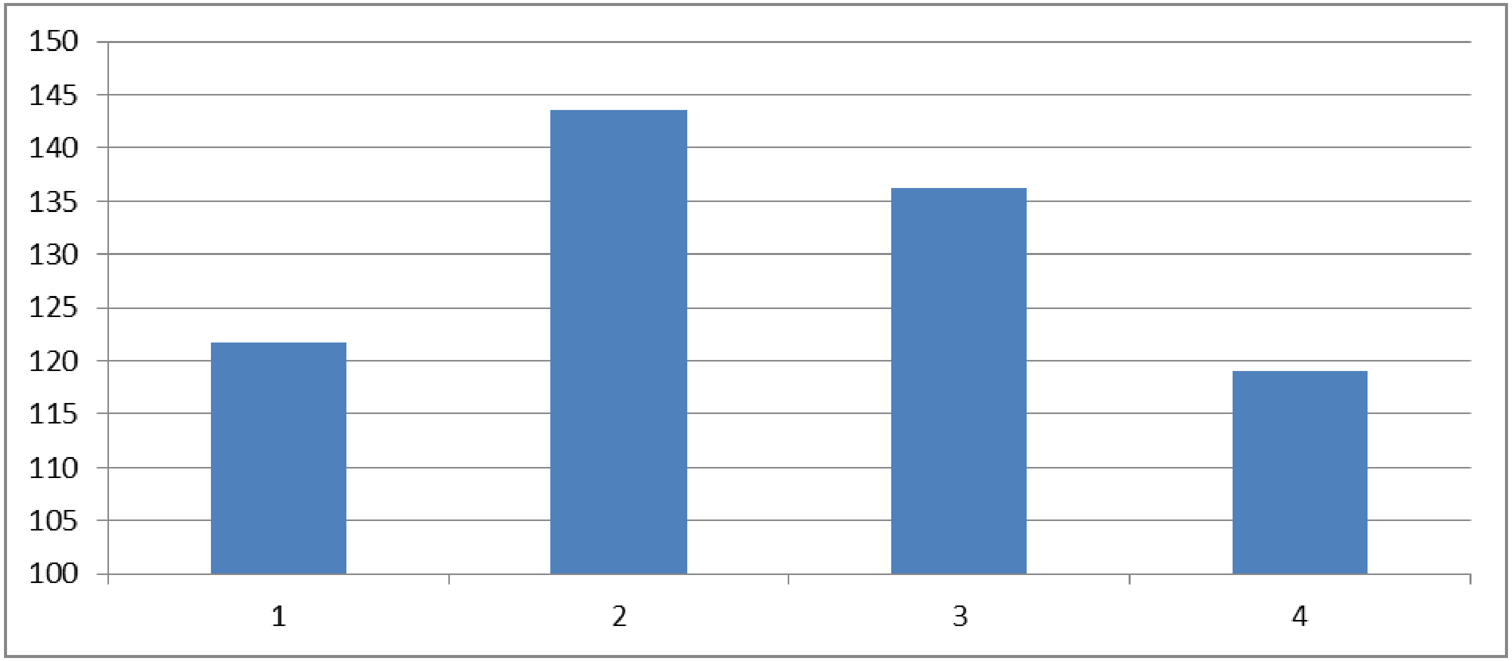


Xinhua/Egypt: Event Types

- Clear that Verbal Cooperation are the most easily forecasted
- They are also rarely spontaneous
- Egypt is major diplomatic power in Middle East, often discussion for week in advance of major summits

Xinhua/Egypt: Num Text Days

- Instead of one day of text, use multiple dates: accuracy best at 2 days, but decreases as noise increases
- Leskovec, Backstrom, and Kleinberg (2009) found that story lines tend to experience their most significant growth/decay cycle over precisely a 48 hour period



Peering Inside Models

- Benefit of using Naïve Baysean model is that we can view the conditional probability table to understand the term biases the model is learning

Xinhua/Egypt: META_GEO (EX1)

- Terms biasing towards 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

Xinhua/Egypt: META_GEO (EX1)

- Terms biasing towards low-event days
- Note "west bank" vs "gaza strip" from the previous table

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

Age of Information

- Does training on an older time period yield worse results? Does the age of information matter?
- Ran three tests, training 2002-2004, 2005-2007, and 2008-2010. For each test, forecasting same 11/2010 – 8/2012 period
- Accuracy changes from +29% to +22% to +25%
- Age of information DOES NOT MATTER

2008-2010 vs 2002-2004 (EX2)

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

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
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

Metanarratives

- Models are learning subnarratives of specific spatial-temporal context and those age over time
- However, subnarratives are rooted in metanarratives that remain constant over this period
- 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 (Egypt as the regional diplomatic superpower)

Other Countries and Sources

- Expanding beyond Egypt:
  - Indonesia: long history of ethnic and religious conflict
  - South Africa: African nation with most coverage and only native English-language country
  - Brazil: largest Latin American country in terms of coverage
  - Germany: 2nd highest European country behind France
- Yields one country in each region, no other datasets exist for these countries other than South Africa

Other Countries and Sources

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

Results

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

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

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

Results

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

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

META_NAME: AFP vs Xinhua (Germany)

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

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

META_NAME/META_GEO Brazil (AFP/Xinhua)

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

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

AFP/Indonesia: Material Conflict

	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

AFP/Brazil: Protest (Code 14XX)

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

Cross-Country Models

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

Weekly Forecasting

- Change to using week's worth of text to forecast events during following week
- Again, forecasting "high event" days
- Up to 45% better than random chance (73% True Pos / 73% True Neg)
- Most predictive is Xinhua coverage of Egypt

CONCLUSIONS

WHAT HAVE WE LEARNED?

A Framework for Forecasting

- The previous sections demonstrate that we have reached a transformative point in the study of human behavior
- A single graduate student can, in the course of a single doctoral dissertation, obtain more than 1.3 billion words of international news coverage covering a quarter-century (with full written authorization) and process that material into 28M events in over 300 categories using just a laptop: no million-dollar projects involving vast teams of highly-trained human analysts

A Framework for Forecasting

- Demonstrates that it is highly feasible to construct fine-grained event datasets for latent forecasting
- Construct ad-hoc physical baselines to test models against

Age of Information & Metanarratives

- Older information is not less predictive than newer information
- Each model captures a specific narrative, but those narratives are wrapped inside of metanarratives that hold over long periods of time
- Can successfully use forecasting as a TOOL to quantitatively capture metanarratives (forecasting as a means, rather than as an ends) – application to digital humanities

Research Question #1

- **What are the latent signatures that 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 highly predictive of future physical behavior. There does not appear to be any one signature or set of signatures that are most predictive, but rather an array of signatures that are closely coupled with the underlying 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. Surprisingly, however, signatures do appear to generalize within a culture over time: while the specific features change, the underlying predictive power of a model does not change as it is trained on older and older information. This suggests the ability of latent models to quantitatively codify the underlying metanarratives that define a culture within the communicative sphere that describes it.

Research Question #3

- **Are signatures universal across classes of physical behavior and intensity levels?** There is strong 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 more considerably more predictable than those with smaller numbers of events.

Forecasting as a Tool

- Here forecasting is used as a tool to capture metanarratives, rather than being the focus
- No optimizations and only basic Naïve Baysean tested
- Sophisticated approaches in literature: Markov-Switching Baysean Vector Autoregressive models, and Latent Dirichlet Allocation models, to Hidden Markov models
- Accuracy around 10-25% better than random chance, right on par with results of this dissertation

Risks & Mitigations

- Dataset Boundaries
- Dataset Errors
- Statistical Limitations
- Adverse Media Effects (Selection bias, media fatigue, helicopter journalism)
- Censorship and Media Interference
  - Blocking Messages
  - Saturating Messages
  - Self-Censorship
- The Paradox of Prediction

The Paradox of Prediction

- Goal of forecasting is to take external action (intervention) to render a correct forecast incorrect
- "Sabotaging the model"

Thank You!

Can We Forecast Conflict?

A Framework for Forecasting Global Human Societal Behavior Using Latent Narrative Indicators