Modeling and Forecasting Violent Intent and Socio-political Contention

Antonio Sanfilippo

Qatar Foundation R&D

Background

- Multi-lab project on Motivation & Intent funded by DHS/S&T, 2004-2009
- Technosocial Predictive Analytics Initiative, Pacific Northwest National Lab, DOE, 2007-2012
- Radical Rhetoric Group, supported,Department of Homeland Security, R&D Directorate (2009-2011)
- Help analysts assess the likelihood of a group to engage in violent behavior
  - Social science encapsulation
  - Content extraction and analysis
  - Modeling and simulation
  - Analytic workflows

VIMS Violent Intent Modeling System

Homeland Security Science and Technology

U.S. DEPARTMENT OF ENERGY

Pacific Northwest NATIONAL LABORATORY

Los Alamos NATIONAL LABORATORY EST. 1943

OAK RIDGE National Laboratory

Sandia National Laboratories

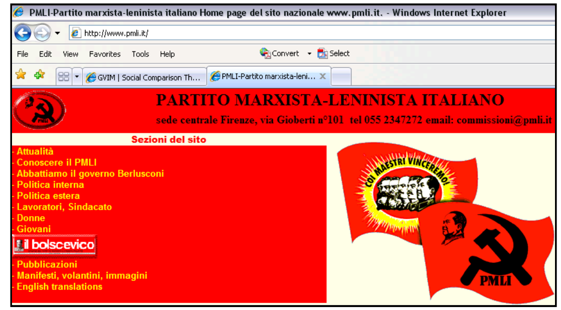
Lawrence Livermore National Laboratory

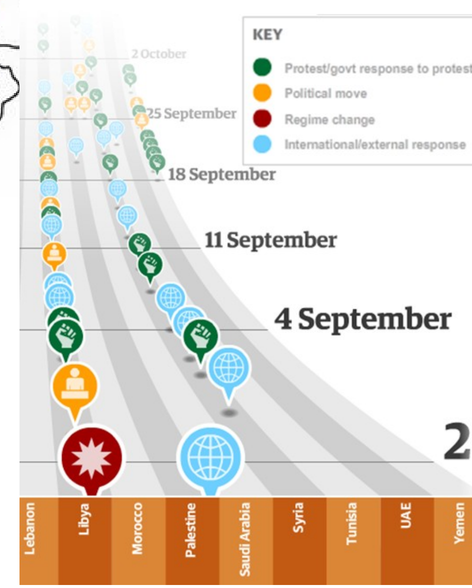
Contributors

- Antonio Sanfilippo (lead)
- Stephen Tratz
- Liam McGrath
- Lyndsey Franklin
- Eric Bell
- Paul Whitney
- Christian Posse
- Bob Baddeley
- Michelle Gregory
- Courtney Corley
- Line Pouchard
- Rick Riesche
- Gary Danielson
- Nick Mileson
- Andrew Cowell
- Amanda Cowell
- Ann Boek
- …

Problem Statement and Approach

- Objective
  - Detect when messages expressing equivalent radical ideologies originate from a terrorist source
- Approach
  - Quantify the co-expression of rhetoric and action to train classification models of violent intent
- Applications
  - Recognize messages from terrorist sources
  - Detect and forecast sociopolitical contention in social media





Developing a scheme to annotate violent intent

- **Framing**
  - How a communication source uses messaging to influence the target audience – *collective action frames*
  - How the target audience responds – *frame resonance*
- **Issues**
  - military, religion, law, security, politics, …
- **Violence Indicators**
  - *Moral disengagement*
  - *Violation of sacred values*
  - *Social isolation*
  - *Violence and contention*

Theories of collective action frames

Gamson

- ***Injustice:*** identify individuals or institution to blame for grievances
- ***Identity***: specify aggrieved group with reference to shared interests and values
- ***Agency***: recognize that grieving conditions can be changed through activism

Snow and Benford

- ***Diagnostic frame***: tell new recruits what is wrong and why
- ***Prognostic frame***: present a solution to the diagnosed problem
- ***Motivational frame***: give people a reason to join collective action

Entman

Substantive frame functions

- Defining effects or conditions as problematic
- Identifying causes
- Conveying moral judgment
- Endorsing remedies or improvements

Substantive frame foci

- Political events
- Issues
- Actors

Frame annotation guidelines

- Formalize frames as **speech acts**
  - Utterances that have performative function in language and communication, e.g. *promise, order, warn* (Austin 1962, Searle 1969)
- A frame is a performative utterance that
  - identifies a **PROMOTER**
  - conveys a particular **INTENTION** in making the utterance
  - may identify a **TARGET**
  - specifies one or more **ISSUES**

Annotation scheme implements "Intelligent union" approach

The Parliamentary Bloc of the Muslim Brotherhood (MB) denounces the insistence of the security apparatus on terrorizing innocent people and on using the emergency law against honest Egyptian citizens, through its campaign of raids and detentions against Muslim Brothers in the governorates of Cairo, Alexandria, Daqahliya and lastly Minya.

PROMOTER

INTENTION

TARGET

ISSUES

- POLITICS
- SOCIAL
- LAW
- SECURITY

PROMOTER

- used by Snow and Benford
- corresponds to the result of Gamson's *identity* frame function
- overlaps with Entman's notion of *actors*

COMMUNICATIVE INTENT

- implicit in the frame classification of Gamson (*injustice, identity, agency*) and Snow and Benford (*diagnostic, prognostic, motivational*)

TARGET

- corresponds to the result of Gamson's *injustice* frame function

ISSUES

- as in Entman

Annotation methodology: promote objectivity and enable automation

- **INTENT** is broken down into 14 speech act classes
  - ASSERT, BELIEVE, CRITICIZE, EXPLAIN, REQUEST, …
  - Each “intention” class has various lexical realizations (from WordNet)

INTENT	CRITICIZE
Lexical realizations	accuse, blame, calumniate, charge, condemn, criticize, denigrate, deplore, impeach, incriminate, lambast, malign, reproach, slander, …

- W
  - ECONOMY, POLITICS, SOCIAL, LAW, MILITARY, ADMINISTRATION, ENVIRONMENT, SECURITY, RELIGION (from WordNet Domains)

Use kappa test to validate annotation: 30 documents with 4 annotators

Cohen kappa test: Human vs. Human (average = 0.70)					
Ratings	**Subject A**	**Subject B**	**Kappa**	**p-value**	**z-score**
1660	1	2	0.783	~0	31.9
1599	1	3	0.928	~0	37.1
1753	1	4	0.553	~0	23.2
1656	2	3	0.809	~0	33
1776	2	4	0.543	~0	22.9
1755	4	3	0.573	~0	24

Fleiss kappa test: Groups of 4 Human Annotators			
Ratings	**Kappa**	**p-value**	**z-score**
1660	0.499	~0	46.2

Linguistic indicators of violent intent

- **Moral disengagement1** (*hate, fear, judge, criticize*)
  - People engage in inhumane conduct to achieve a goal believed to be morally right
  - Removal of ethical restrictions against violence through acts of dehumanization
- **Violation of sacred values2,3** (*military, religion*)
  - Ideals of love, honor, justice and religion come under secular assault and people struggle to protect themselves from moral contamination
- **Social isolation4** (*confine, abandon, withdraw*)
  - Requirement that a recruit cut off ties to family, friends, and anyone else outside the organization
- **Violence and contention** (*attack, fight, kill*)

1*Badura 1999;* **2,3***Stenberg 2003, Rice 2009; Tetlock et al. 2000;* **4***Navarro 2009.*

Violent intent annotation scheme

160 categories covering some 13,000 word meanings

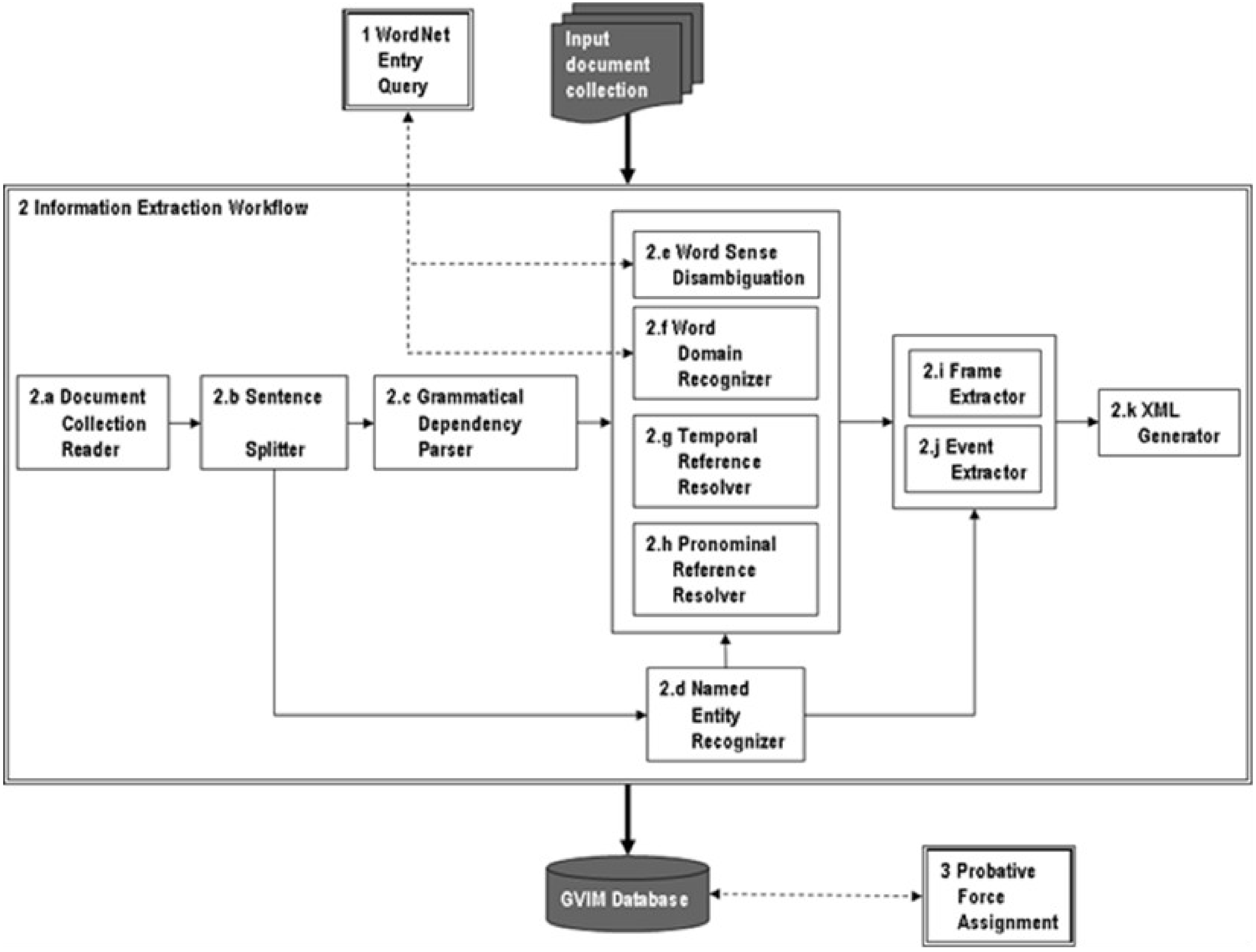
Using text mining to automate violent intent annotation

Egypt Bureau of Democracy, Human Rights, and Labor, February 28, 2005

… In *March 18, 2004* *Abdelwahab Ads*, deputy editor of Al Jumhuriya, *accused* the *Jews* of the terrorist attack in Madrid on March 11. …

PROMOTER	=	*Abdelwahab Ads*
C-INTENT:CRITICIZE	=	accused
TARGET	=	the Jews
ISSUES	=	[SECURITY=0.50, POLITICS=0.50]
EVENT_DATE	=	[YEAR=2004, MONTH=03, DAY=18]
PUBLISH_DATE	=	[YEAR=2005, MONTH=02, DAY=28]

NLP Pipeline



Evaluate automatic frame extraction

- Used kappa and precision/recall tests to evaluate of manually and automatically assigned annotations to 30 documents

Cohen kappa test: Human vs. Computer (average = 0.52)					
Ratings	**Subject A**	**Subject B**	**Kappa**	**P-value**	**z-score**
1625	1	Computer	0.605	~0	25.2
1683	2	Computer	0.507	~0	21.7
1626	3	Computer	0.613	~0	25.6
1763	4	Computer	0.339	~0	14.8

Fleiss kappa test: 4 Human Annotators plus Computer			
Ratings	**Kappa**	**p-value**	**z-score**
1433	0.422	~0	50.5

#Correct	#Incorrect	Precision	Recall	F1
157	43	0.785	0.698	0.739

(frame detection only)

Modeling violent intent: A data driven approach

- Data*

	Terrorist groups	Non-Terrorist groups
Regional	*al Qa'ida in the Arabian Peninsula*	*Movement for Islamic Reform in Arabia*
Transnational	*al Qa'ida Central*	*Hizb ut-Tahrir* ("Party of Liberation")

- ...cuments with violent intent features automatically
- Learn classification model from annotations that recognize documents from terrorist and non-terrorist sources
  - Identify contributing features and their relative weight

Provided by DHS/HFD, see Smith et al. 2008

Used multinomial Bayesian classification* to model of radical rhetoric

- The probability that a document *D* belongs to a class *C* is

$$p(C|D) = \frac{p(D|C) * p(C)}{p(D)}$$

Where $p(D)$ is the probability of generating a document that has D's features

- The probability of a document given its class is derived as the product of the probabilities of the features occurring in the document

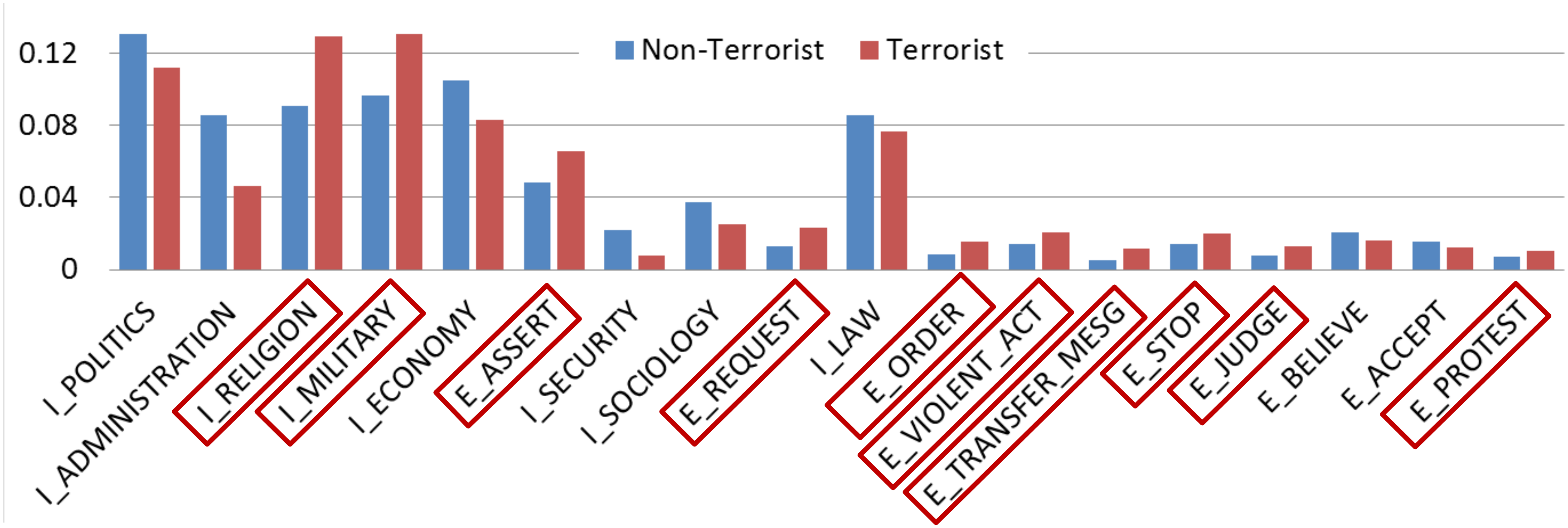
$$p(D|C) = N! * \prod_{i=1}^{k} \frac{P_i^{n_i}}{n_1!}$$

- N is the number of features in *D*
- $n_1, n_2, \ldots, n_k$ is the number of times feature *i* occurs in document *D*
- P_i is the probability of obtaining the feature *i* from documents of class *C*.

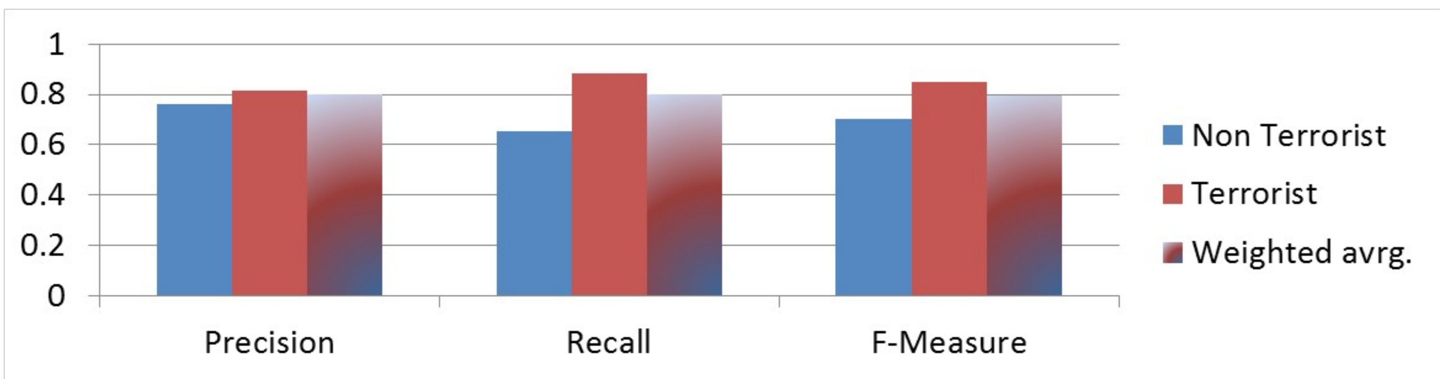
*McCallum & Nigam, 1998

Modeling results and evaluation

- Detecting violent & non-violent radical rhetoric, top 18 factors



- Evaluation results



Framing contention in Twitter data

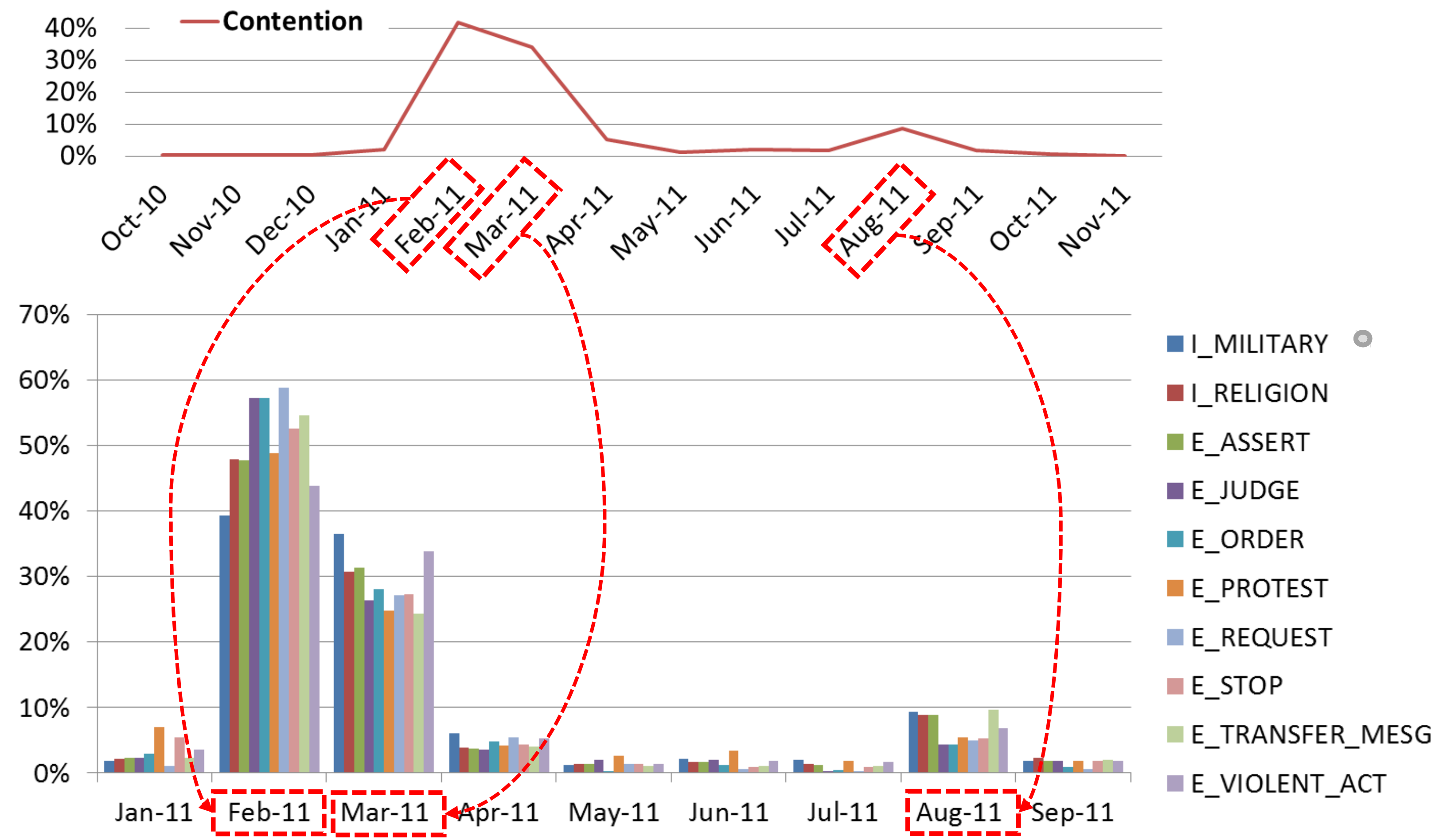
- Harvested Twitter postings about Syria, Egypt, Tunisia and Libya for the period related to the Arab Spring
- Processed Twitter postings using the Frame Analysis platform
- Measured the occurrence of top frame features highly correlated with terrorist rhetoric to assess sociopolitical contention

$$contention\left(message_j\right) = \sum_{i=1}^{n} |F_{ij}| * p(F_i)$$

  - $|F_{ij}|$ is the number of times a feature F_i occurs in message *j*
  - $p(F_i)$ is the probability with which the feature F_i identifies terrorist rhetoric in the referent model

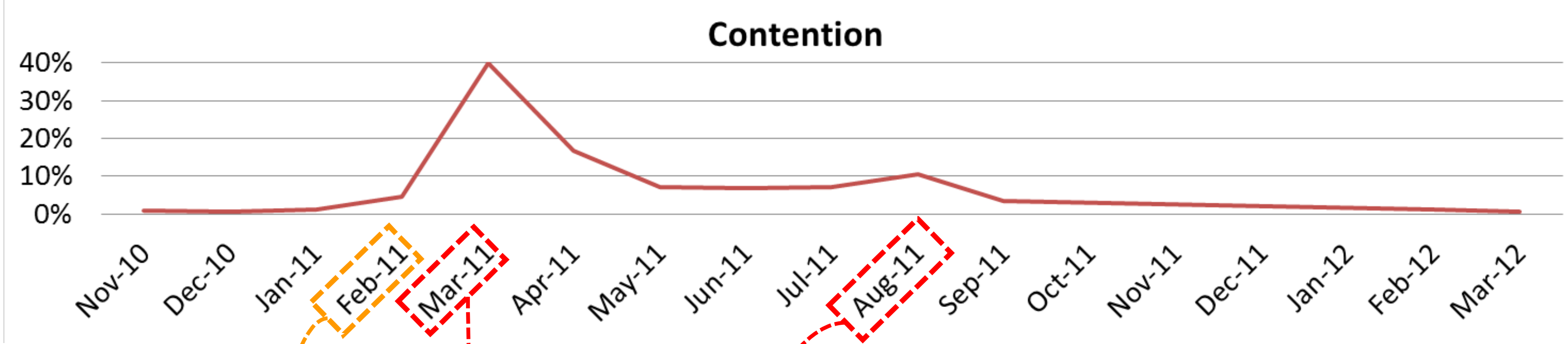
By Area: Libya

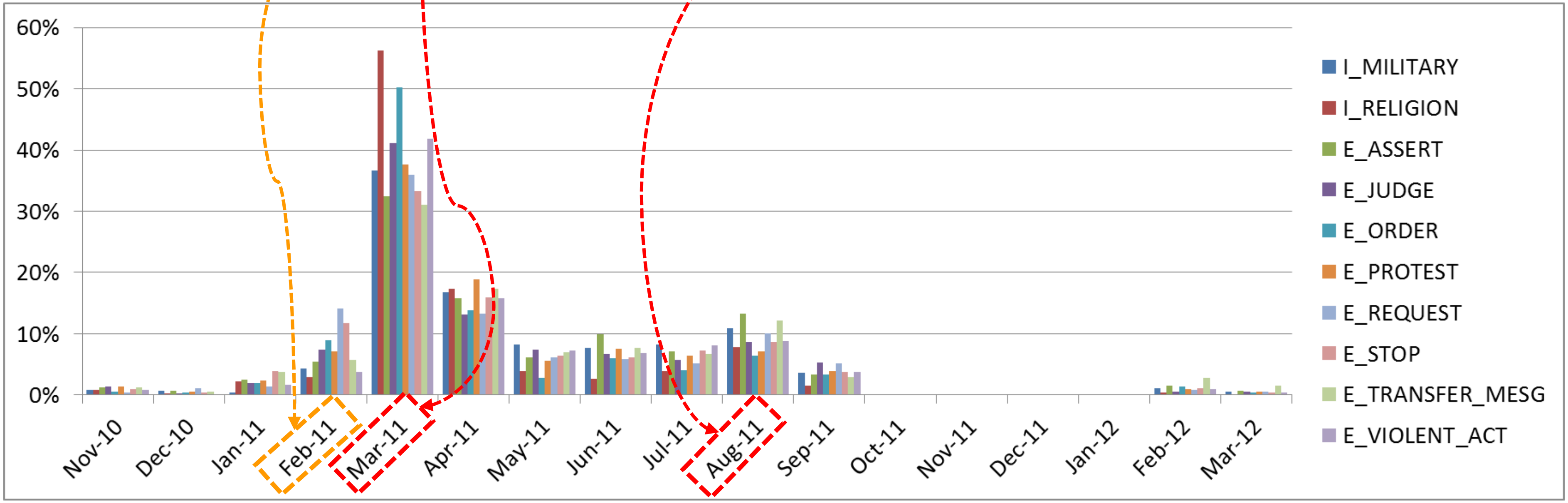
- 2/16/11 Protests start in Benghazi
- 3/5/11: Counterattack by Gaddafi
- 8/26/22 Rebels move interim government to Tripoli



By Area: Syria

- 3/19/2011: Syrian security forces kill protesters
- 4/8/2011: 109 people killed in Hamas



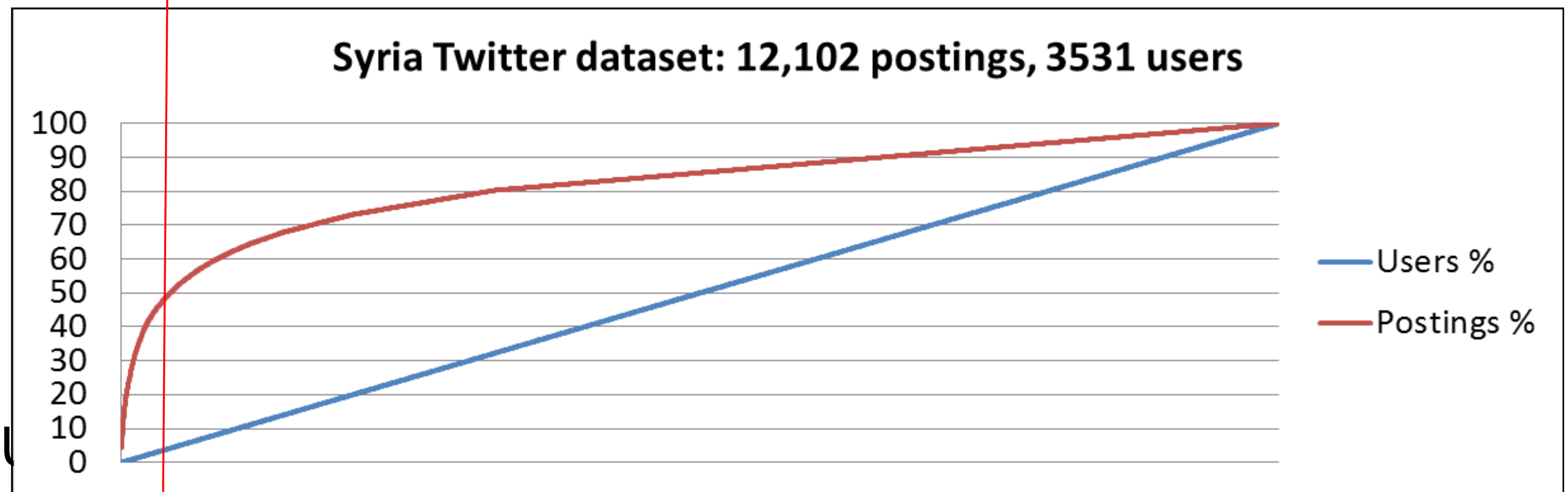


Discussion: Factors promoting contention

- religion and military together are the most influential features in characterizing contention in Arab Spring tweets
  - *Violation of sacred values* as the main overall factor
- After the initial outbreak of contention, religion tends to decrease more quickly than military
- assertion is the leading feature of communicative intent
- Caveats
  - Violent intent annotation was developed for longer documents
  - Only English Twitter content was processed

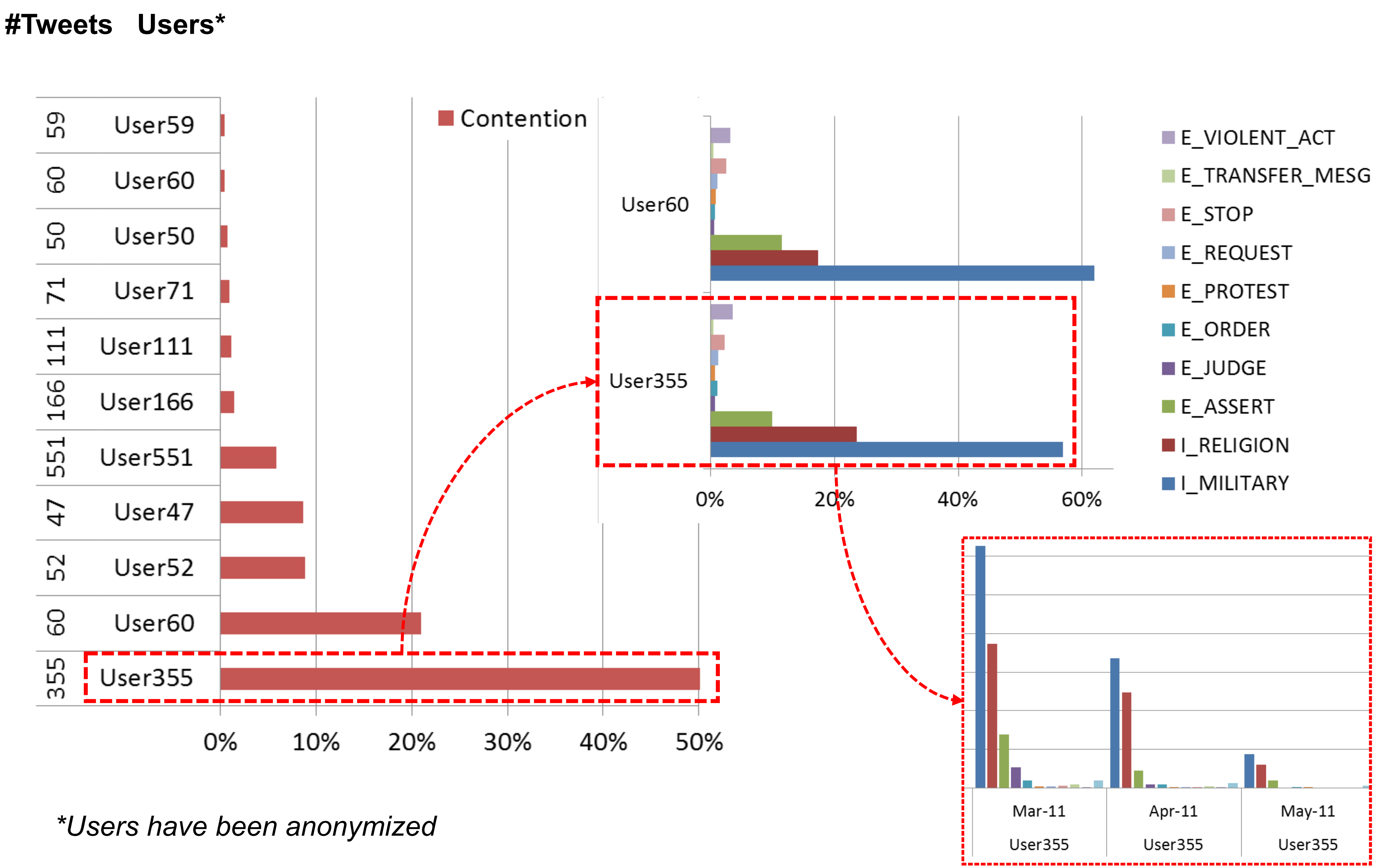
Differentiating High and Low Users

- On average, about 50% of postings are generated by <5% of users, each with >10 postings



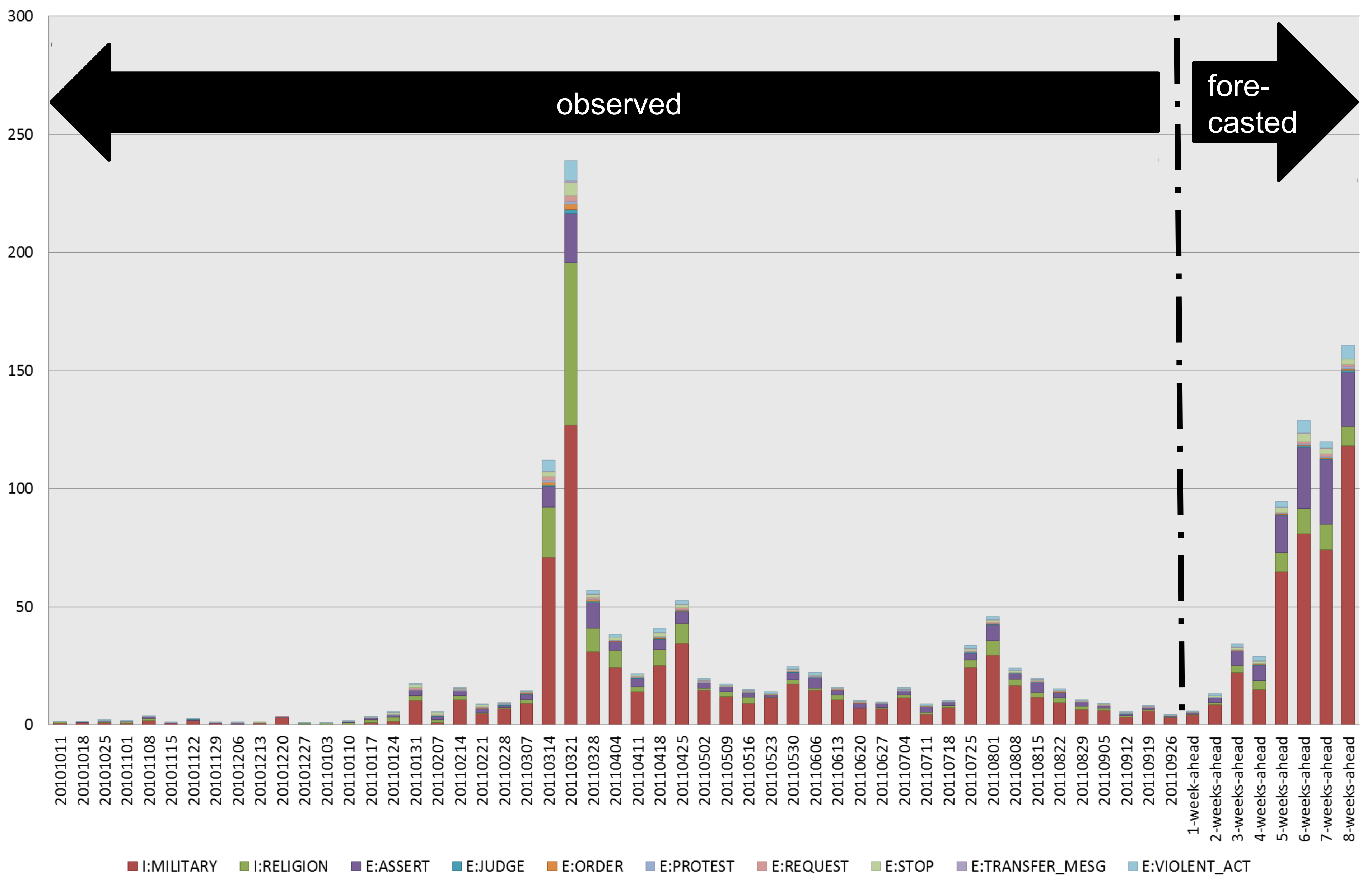
- U
discussion threads and influence others: *trend initiators/setters*
- Users with fewer postings tend to be followers: *trend adopters/consumers*

Framing contentious rhetoric of high users (Syria postings)



Users have been anonymized

Forecasting Socio-Political Contention

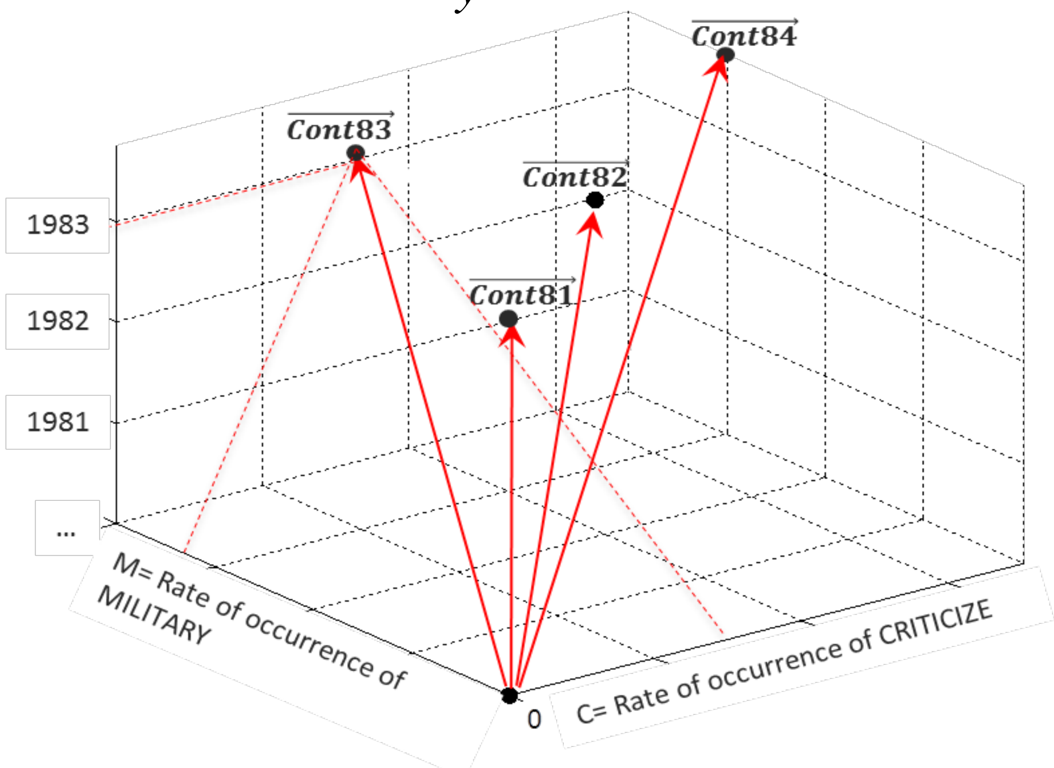


Multivariate Time Series Forecasting with Support Vector Regression*: *Training*

- Each training sample is a pair $\{\vec{x}, y\}$, where $\vec{x}$ is a vector for the time-series class to be learned, and y the associated set of values

$$\{\underbrace{\overrightarrow{Cont83} = [\overrightarrow{Cont82}, \overrightarrow{Cont81}, \ldots]}_{\vec{x}}, \underbrace{[M=0.3, C=0.1, \ldots]}_{y}\}$$

  - *Cont* = "contention"
  - M = MILITARY
  - C = CRITICIZE



*Smola et al. (1997). We used the Weka implementation (Pentaho).

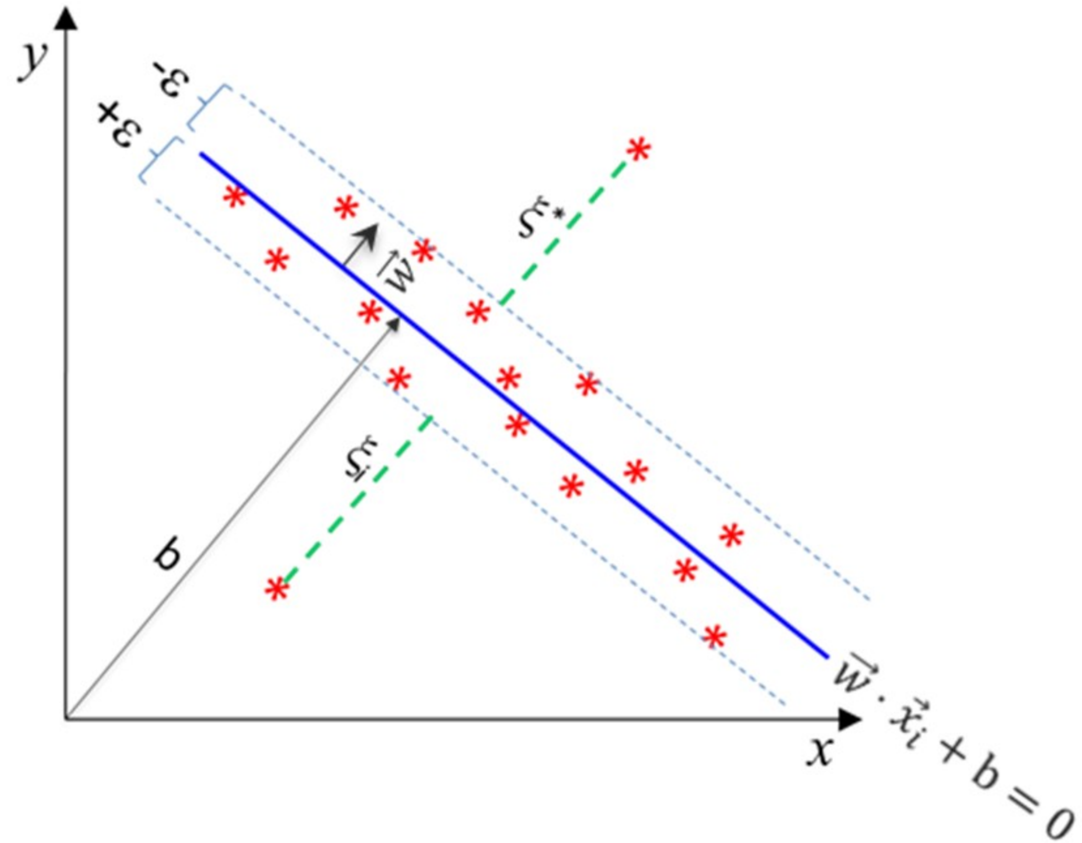
Time Series Forecasting with Support Vector Regression: *Regression*

- Find a function that for each vector $\vec{x}_i$ in the training dataset approximates its set of values y_i within ε-deviation with no penalty, and within ξ-deviation with increasing penalty:

$$y_i = \vec{w} \cdot \vec{x}_i + \mathrm{b}$$

$$\text{for } i = 1, \dots, n$$

  - Minimize the length of weight vector $\vec{w}$ to avoid over-fitting

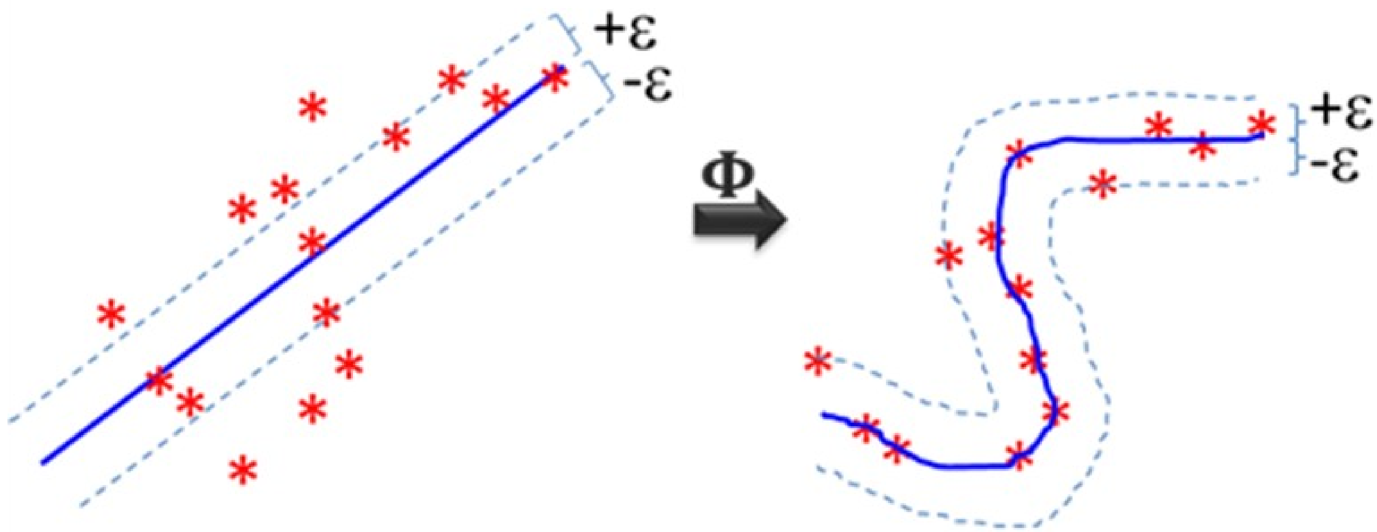


Time Series Forecasting with Support Vector Regression: *Non-linearity*

- Map the vector data into a multidimensional feature space using a kernel function Φ to deal with non-linear problems

$$y_i = \Phi(\vec{w}) \cdot \Phi(\vec{x}_i) + \mathrm{b}$$

e.g. using the polynomial kernel: $\Phi(\vec{w}) \cdot \Phi(\vec{x}_i) = (1+\vec{w} \cdot \vec{x}_i)^3$

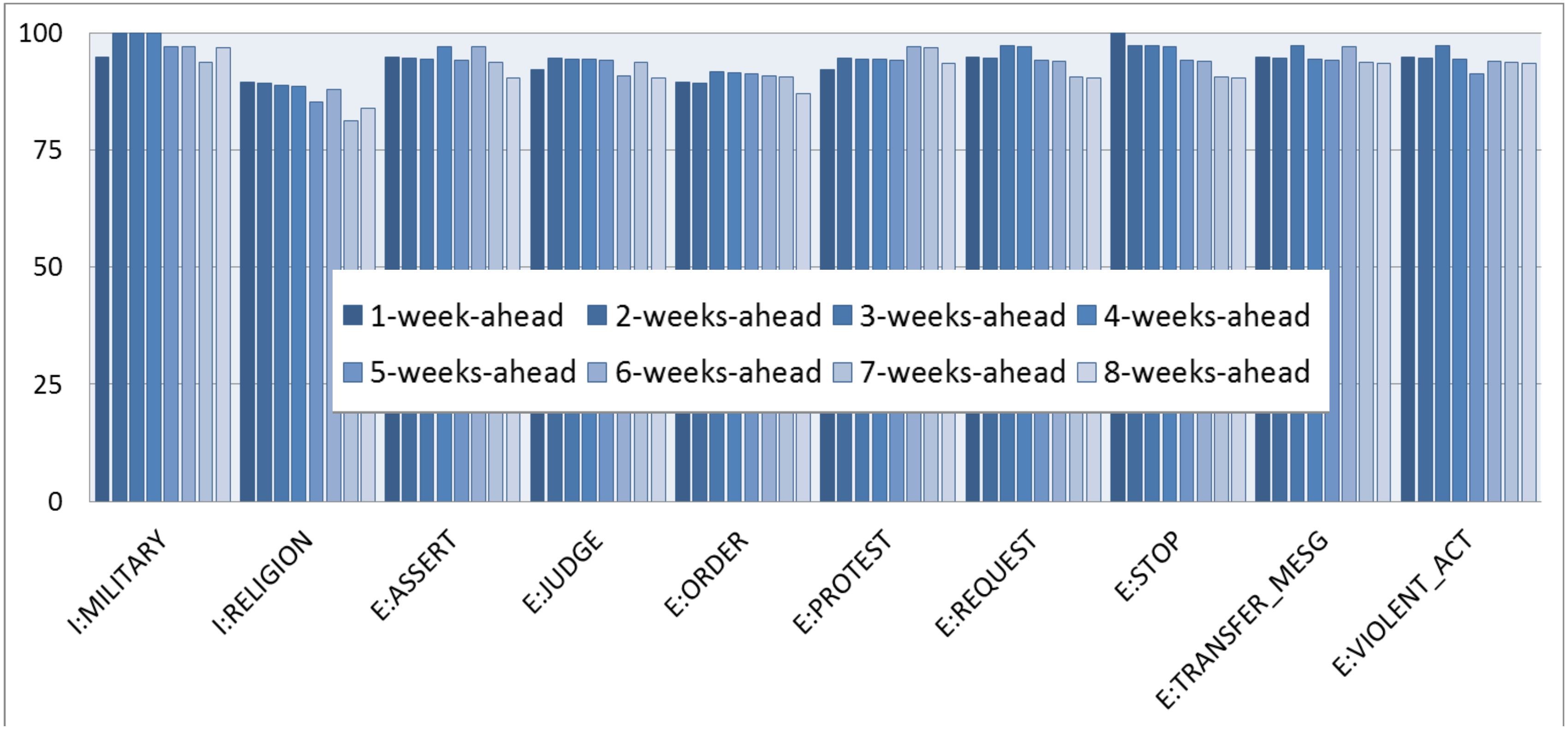


Evaluating forecasting results using direction accuracy

Direction accuracy = $\frac{1}{n}\sum_{i=1}^{n}\alpha_i$

- where: $\alpha_i = 1, if\ sgn(current_i - previous_i) = sgn(predicted\_current_i^* - predicted\_previous_i^*)$

  $\alpha_i =$ 0 otherwise,



Related Work*

- Manual coding
  - Content categories are defined as sets of words based on explicit rules of coding, e.g. Integrative Complexity
  - Smith et al. (2008), Winter (2011), Suedfeld & Brcic (2011), and Conway et al. (2011)
  - Assessments based on subject matter experts' answers to questionnaires
  - Borum et al., 2006; Webster et al., 2000; Pressman, 2009
- Automated coding
  - Exploit the difference between content and function words (Pennebaker 2011)
  - Use text mining techniques to extracting sociocultural and psychosocial signatures as specified by theoretical approaches to social and political analysis
  - *Leadership Traits Analysis* (Hermann and Sakiev 2011)
  - *Operational Code* detailing values and world views of political actors (Walker 2011)

Dynamics of Asymmetric Conflict, 4(2), 2011.

Conclusions

- Modeling radical rhetoric to identify violent intent helps detect messages from terrorist sources
- The ensuing models can be applied to social media data to “take the pulse” of social contention through time
- The emerging time series data can be use to make forecasts using time series modeling techniques

Than ks!

Probability of a document given its class in MBN – a simple example

- Only two features in our category scheme: I_RELIGION, I_LAW
- Assume that in the training documents
  - $p(\text{I\_RELIGION}|\, terrorist) = 0.75$
  - $p(\text{I\_LAW}|\, terrorist) = 0.25$
- Document *D* has only 3 annotations
  - {I_LAW, I_RELIGION, I_LAW}
- According to the formula

$$p(D|C) = N! * \prod_{i=1}^{k} \frac{P_i^{n_i}}{n_1!},$$

- the probability of D given the *terrorist* class is computed as:

$$p(\{\text{I\_LAW, I\_RELIGION, I\_LAW}\}|\, terrorist) = 3! * \frac{0.25^2}{2!} * \frac{0.75}{1!} = 6 * \frac{0.06}{2} * \frac{0.75}{1} = \frac{9}{64} = 0.12$$

Support Vector Regression*

- Find $y_i = \vec{w} \cdot \vec{x}_i + \text{b}$ that has at most ε deviation from $y_{1,\ldots,n}$ for $\vec{x}_{1,\ldots,n}$ and is as “flat” as possible (to avoid over-fitting)
  - Minimize the length of the weight vector ($\|\vec{w}\|$) using penalty variable ξ:

$$\min \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i{}^*)$$

subject to:

$$y_i - (\vec{w} \cdot \vec{x}_i + b) \leq \varepsilon + \xi_i$$

or

$$y_i - (\vec{w} \cdot \vec{x}_i + b) \geq -\varepsilon - \xi_i^*$$

for $i = 1, \ldots, n$