The Hard Problem of Prediction for Prevention

Reading Between the Lines

Hannes Mueller IAE (CSIC)
Christopher Rauh (University of Montreal

Fundació Economia Analítica

April 2018

Financial Support from Fundación BBVA

Introduction

- Civil wars are a serious humanitarian and economic problem.
- And we fail to prevent them.
- This is reflected in large expenditures on crisis response.
  - Humanitarian response: ca. 24.5 billion US dollars in 2014
  - Peacekeeping: ca. 8 billion US dollars per year.
- Review of the United Nations Peacebuilding Architecture (2015):
  - *If more global priority were consistently given to efforts at sustaining peace, might there not, over the course of time, be reduced need for crisis response?*

This Talk

- Define hard problem:
  - why *hard*
  - why *problem*
- Literature
- Solution:
  - topic model with news text $\longrightarrow$ summaries of text
  - use summaries to get early warning
  - speculate why it works

Data/Definitions

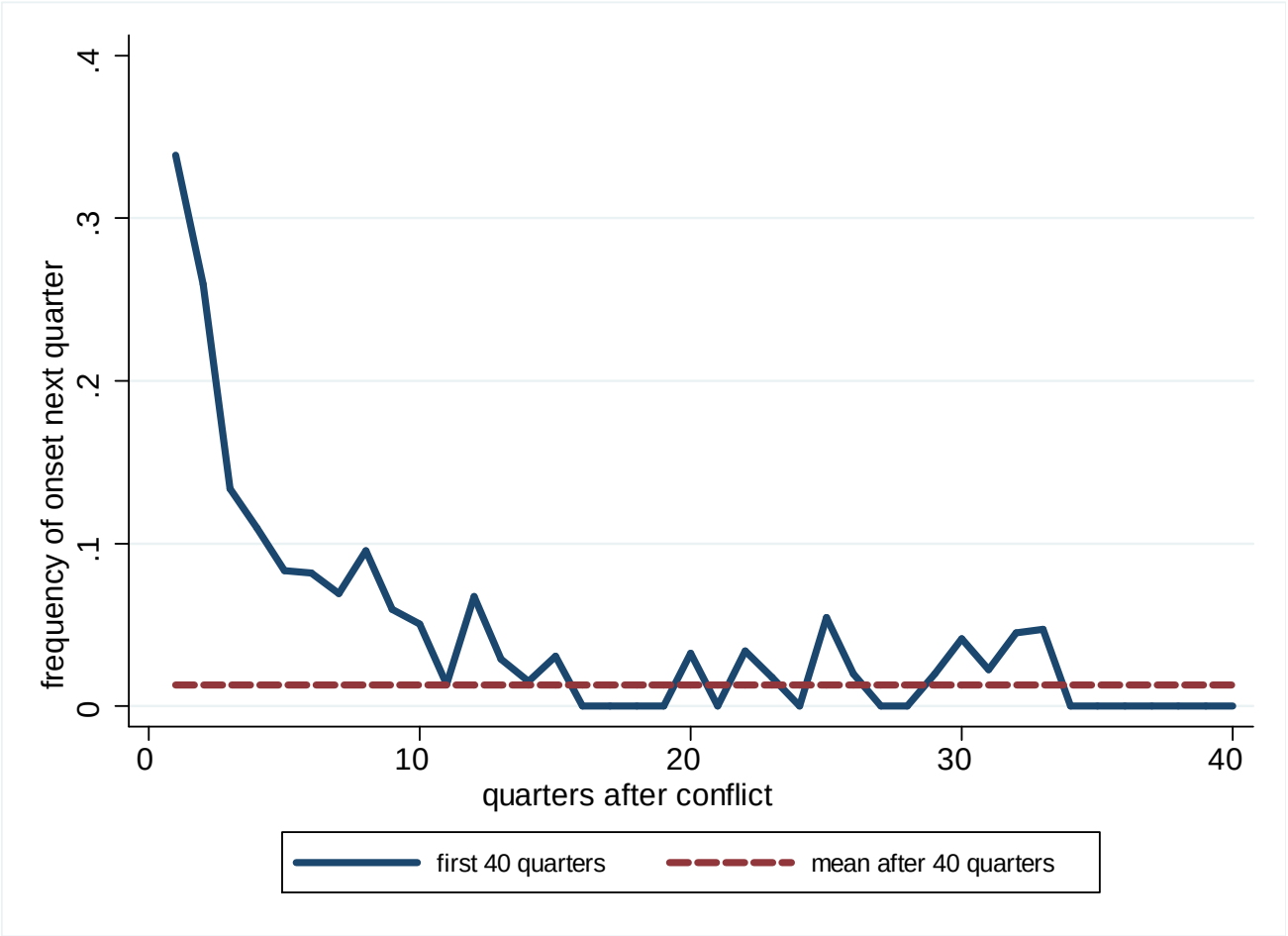
- Violence data from Uppsala Conflict Data Program (UCDP)
- Georeferenced Event Dataset (GED), Sundberg and Melander (2013)
- Gives quarterly data for all countries 1989-2016
- Internal conflicts: state-based conflict, non-state conflict, one-sided violence.
- We focus on onset of conflict, i.e. code a dummy of start.
- Conflict in the literature is defined as 25+ or 1000+ battle-related deaths per year.
- Not obvious how to translate this to the quarterly data.
- We use three thresholds: 1, 50 and 500 (all violence, top 50% and top 25%)

The Hard Problem of Prediction

- Take the 50+ definition.
- There were 433 onsets in almost 19,000 observations.
- Conflict history is a strong predictor of conflict onset.
- 359 onsets followed within 10 years of another conflict, 75 were "new" onsets.
- The following plot shows the risk of an onset post-conflict.

The Hard Problem of Prediction

Onset likelihood in post-conflict period (50+):



The Hard Problem of Prediction

- Average risk outside the 10-year period is very low: about 0.6 percent.
- This means post-conflict period is a powerful forecast.
- After around 40 quarters, however, risk is again close to "normal".
- Call the 40 quarters after conflict post-conflict.
- Call all other quarters with peace pre-conflict.

The Hard Problem of Prediction

- Then we can write down a simple Markov transition matrix:

		this quarter		
		peace pre-conflict	conflict	peace post-conflict
next quarter	peace pre-conflict	99.45%	0.00%	1.45%
	conflict	0.55%	80.21%	8.26%
	peace post-conflict	0.00%	19.79%	90.29%

The Hard Problem of Prediction

- Peace always follows peace but is much less stable post-conflict.:

		this quarter		
		peace pre-conflict	conflict	peace post-conflict
next quarter	peace pre-conflict	99.45%	0.00%	1.45%
	conflict	0.55%	80.21%	8.26%
	peace post-conflict	0.00%	19.79%	90.29%

- Still, we have 75 onsets pre-conflict.
- This is what we call the *hard problem*.

The Hard Problem: Why Should We Care?

- Ironically, because it is hard.
- Pre-conflict peace is stable.
- Post-conflict peace is unstable.
- After 30 quarters, the distribution when starting from the pre-war peace is

$$(0.86, 0.05, 0.09)$$

- We have an 86 percent likelihood to be in pre-war peace and a 5 percent chance to be in conflict.
- But when starting from post-war peace the distribution is

$$(0.26, 0.23, 0.51)$$

- We have a 51 percent likelihood to be in post-war peace and a 23 percent chance to be in conflict.

The Hard Problem: Why Should We Care?

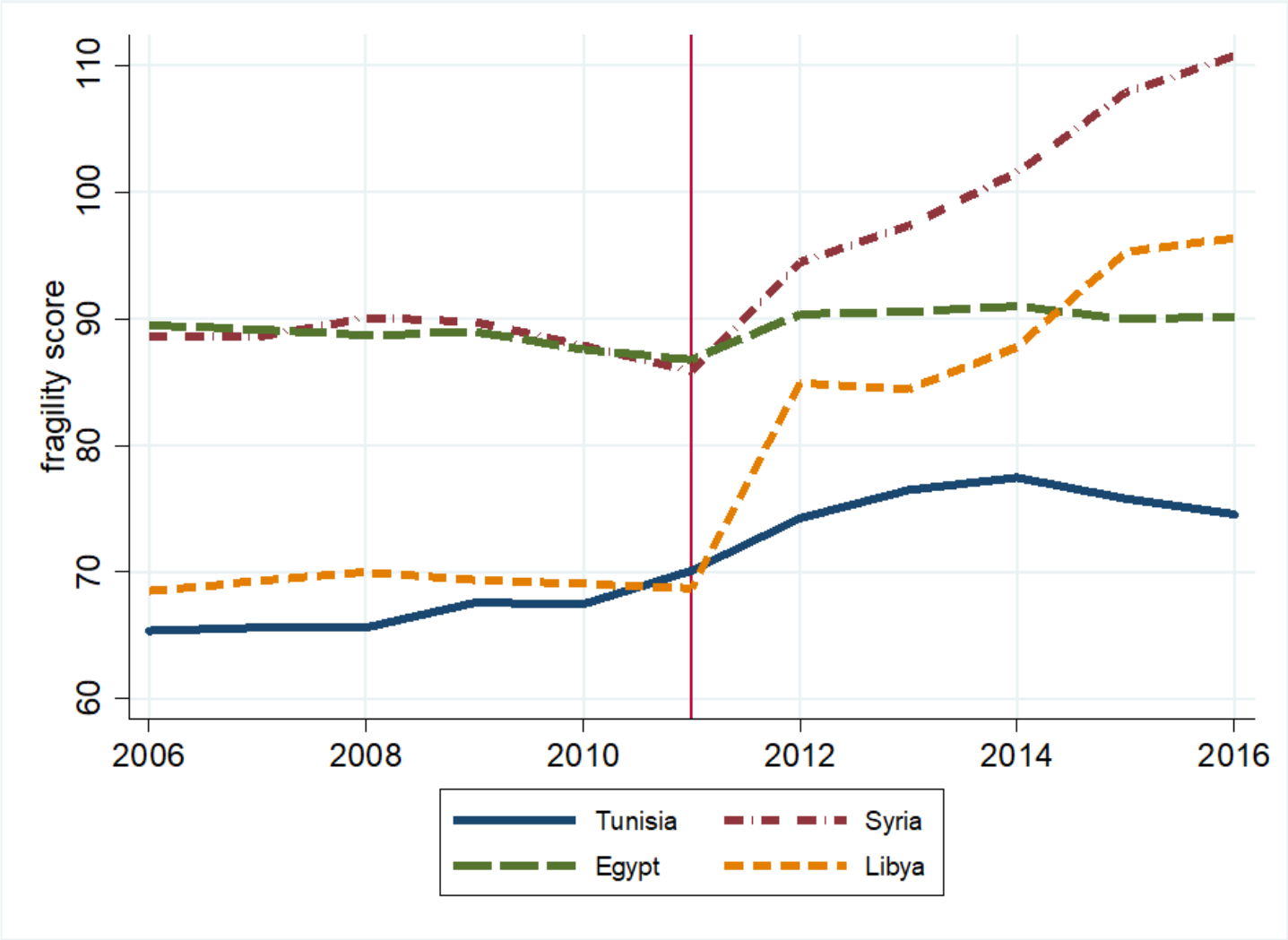
- High risk post-conflict means that it is a bad state to be in.
- If you prevent an escalation into conflict pre-conflict you prevent a country from entering a bad cycle.
- This makes prediction pre-conflict particularly important.

This Talk

- Define hard problem:
  - why *hard*
  - why *problem*
- Literature
- Solution:
  - topic model with news text $\longrightarrow$ summaries of text
  - use summaries to get early warning
  - speculate why it works

Literature (Practise)

Standard fragility measure from the fund for peace.



Literature

- Drivers of conflict:
  - Ethnicity (Montalvo and Reynal-Querol (2005, AER), Esteban et al (2014, AER), Michalopoulos and Papaioannou (2016, AER),
  - Weather (Miguel et al (2004, JPE), Hsiang et al (2013, Science), Ciccone (2011, AEJApplied), Searson (2015, JDE))
  - Commodities (Bazzi and Blattman (2014, AEJMacro), Berman et al (2017, AER))
  - Political Institutions (Besley and Persson (2011, QJE))
- Forecasts:
  - Goldstone et al (2010, AJPS), Chadefaux (2012, JPR),
  - ICEWS, Ward et al (2011)
- Impossibility of perfect forecast: Chadefaux (2017, JCR), Cedermann and Weidmann (2017, Science)

Literature

- Mueller and Rauh (forthcoming, APSR)
- Country fixed effects as a solution for the hard problem.
- For forecast, use only within variation.
- Advantage: you can be sure you predict the timing.
- Disadvantage: you throw away meaningful variation.
- Will try to solve the hard problem differently.

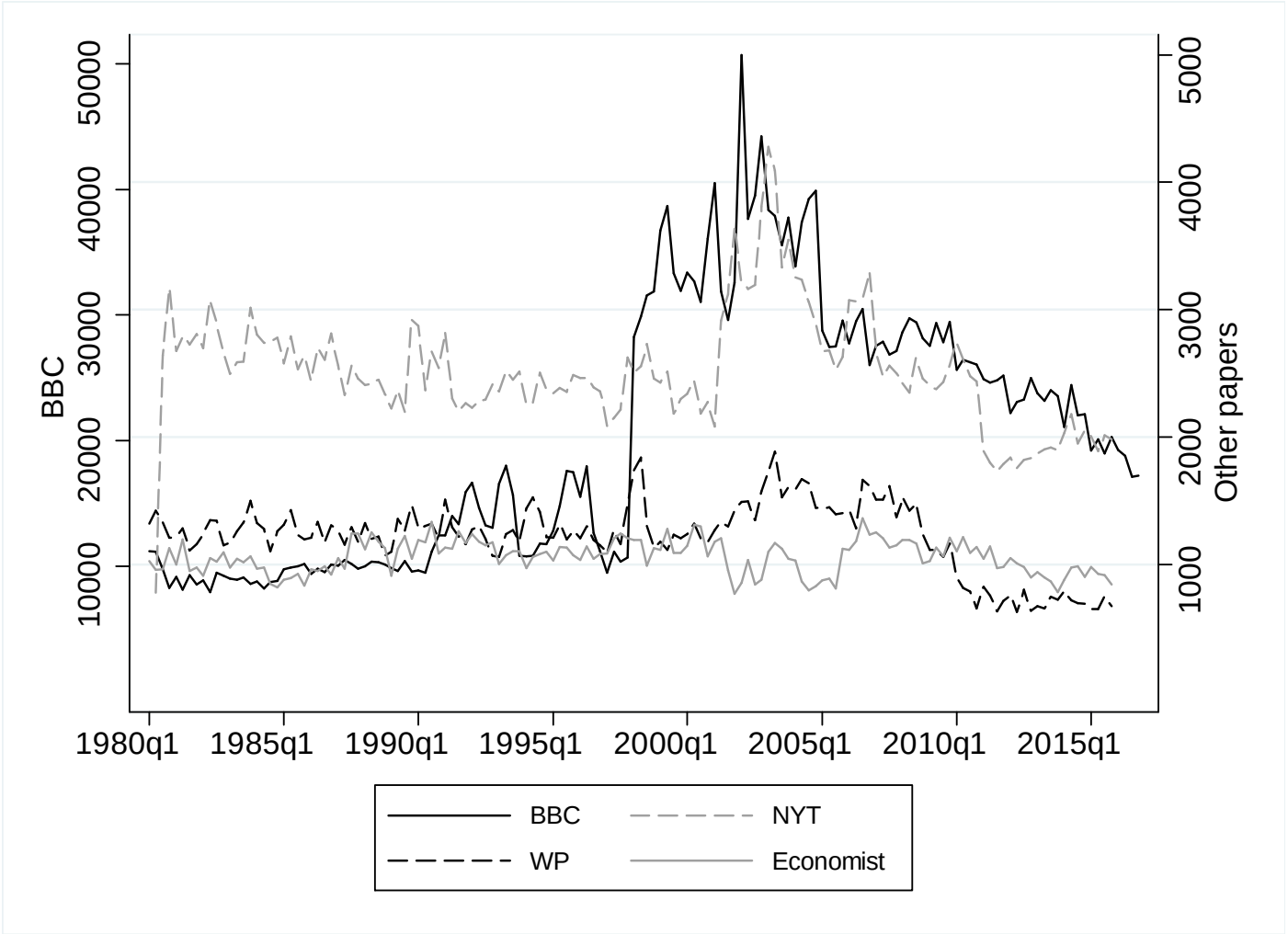
Literature: Machine learning

- Mullainathan and Spiess (2017, JEP)
- Machine learning is great in forecasting: put in x and get out $\hat{y}$
- In economics we are generally interested in hypothesis testing.
- But: useful for heterogenous data (text, images, recording, video).
  - Donaldson and Storeygard on images (2016, JEP)
  - Gentzkow et al on text (2017, NBER)
- We will use it in two ways:
  - unsupervised (feature extraction)
  - supervised (forecast)

Data

- 700,000 articles from New York Times, Washington Post and Economist
- 3.7 million articles from BBC Monitor
- BBC Monitor tracks broadcast, press and social media sources in multiple languages from over 150 countries worldwide.
- Journalists filter, translate and report breaking news.
- We download an article if a country name or capital name is in the title.
- about 4.4 million articles dated from 1989q1 to 2017q3 on over 190 countries.

Number of Articles over Time



Topic Model

- We treat each article m as a vector of tokens $\mathbf{w}_m$ (police, bank, american president, united nations...)
- After dropping rare tokens we have 0.8 million tokens. Need to reduce dimensionality.
- Latent Dirichlet allocation (LDA) introduced by Blei, Ng, and Jordan (2003).
- Topics: probability distributions over the tokens.
- Text generation: journalist picks topic randomly then randomly picks tokens.
- "Latent": only the $\mathbf{w}_m$ are actually observed.
- We tried $K = 5$, 10, 15 topics: low number of topics.
- The following pictures visualize our procedure.

Example: NYT - March 29, 1991. Libya

The exiled Prince Idris of Libya has said he will take control of a dissident Libyan paramilitary force that was originally trained by American intelligence advisers, and he has promised to order it into combat against Col. Muammar el-Qaddafi, the Libyan leader. The United States' two-year effort to destabilize Colonel Qaddafi ended in failure in December, when a Libyan-supplied guerrilla force came to power in Chad, where the original 600 commandos were based. The new Chad Government asked the United States to fly the Libyan dissidents out of the country, beginning a journey that has taken them to Nigeria, Zaire and finally Kenya. So far, no country has agreed to take them permanently. The 400 remaining commandos, who have been disarmed, were originally members of the Libyan Army captured by Chad in border fighting in 1988. They volunteered for the force as a way of escaping P.O.W. camps. "Having received pledges of allegiance from leaders of the force, Prince Idris has stepped in to assume responsibility for the troops' welfare," said a statement released in Rome by the royalist Libyan government in exile. It was overthrown in 1969.

Example: NYT - March 29, 1991. Libya (Stopwords)

the exiled prince idris of libya has said he will take control of a dissident libyan paramilitary force that was originally trained by american intelligence advisers, and he has promised to order it into combat against col. muammar el-qaddafi, the libyan leader. the united states' two-year effort to destabilize colonel qaddafi ended in failure in december, when a libyan-supplied guerrilla force came to power in chad, where the original 600 commandos were based. the new chad government asked the united states to fly the libyan dissidents out of the country, beginning a journey that has taken them to nigeria, zaire and finally kenya. so far, no country has agreed to take them permanently. the 400 remaining commandos, who have been disarmed, were originally members of the libyan army captured by chad in border fighting in 1988. they volunteered for the force as a way of escaping p.o.w. camps. "having received pledges of allegiance from leaders of the force, prince idris has stepped in to assume responsibility for the troops' welfare," said a statement released in rome by the royalist libyan government in exile. it was overthrown in 1969.

Example: NYT - March 29, 1991. Libya

exiled prince idris libya control dissident libyan paramilitary force originally trained american intelligence advisers, promised order combat col. muammar el-qaddafi, libyan leader. united states' two-year effort destabilize colonel qaddafi ended failure december, libyan-supplied guerrilla force came power chad, original 600 commandos based. new chad government asked united states fly libyan dissidents country, beginning journey taken nigeria, zaire finally kenya. far, country agreed permanently. 400 remaining commandos, disarmed, originally members libyan army captured chad border fighting 1988. volunteered force way escaping p.o.w. camps. "having received pledges allegiance leaders force, prince idris stepped assume responsibility troops' welfare," statement released rome royalist libyan government exile. overthrown 1969.

Example: NYT - March 29, 1991. Libya (Lemmatizing)

exiled prince idris libya control dissident libyan paramilitary force originally trained american intelligence advisers, promised order combat col. muammar el-qaddafi, libyan leader. unit state two-year effort destabilize colonel qaddafi ended failure december, libyan-supplied guerrilla forces came power chad, origin 600 commando based. new chad government asked united states fly libyan dissidents country, beginning journey taken nigeria, zaire finally kenya. far, country agreed permanently. 400 remain commandos, disarmed, originally members libyan army captured chad border fighting 1988. volunteered force way escaping p.o.w. camps. "having received pledges allegiance leader force, prince idris steped assume responsibility troop's welfare," statement released rome royalist libyan governmeant exile. overthrown 1969.

Example: NYT - March 29, 1991. Libya

exil princ idri libya control dissid libyan paramilitari forc origin train american intellig advisers, promis order combat col. muammar el-qaddafi, libyan leader. unit state two-year effort destabil colonel qaddafi end failur december, libyan-suppli guerrilla forc came power chad, origin 600 commando based. new chad govern ask unit state fli libyan dissid country, begin journey taken nigeria, zair final kenya. far, countri agre permanently. 400 remain commandos, disarmed, origin member libyan armi captur chad border fight 1988. volunt forc way escap p.o.w. camps. "have receiv pledg allegi leader force, princ idri step assum respons troop welfare," statement releas rome royalist libyan govern exile. overthrown 1969.

Make Topics with Gibbs Sampler

- Feed 4.4 million texts, $\mathbf{w}_m$, into algorithm.
- https://radimrehurek.com/gensim/models/ldamodel.html
- Approximates LDA model to back out
  - probability distribution over tokens for K topics
  - share of each topic in each text, $\boldsymbol{\eta}_m$
- Use of co-occurrance to build topics.
- Reduces 0.8 million token counts to 5, 10 or 15 shares.

intern
import
new
relat
issu
develop
visit
meet
trade
beij
econom
offici
reform
secur
deleg
headlin
chang
ministri
establish
member
term
open
polit
progress
level
support
high
affair
concern
partner
central
process
promot
law
stabil
initi
increas
expect
tie
note
sign
hope
parti
good
institut
union
european
improv
bilater
talk
global
agre
time
make
discuss
way
attend
problem
strengthen
economi
includ
diplomat
bank
leader
websit
work
opportun
strateg
role
sector
peac
project
ambassador
invest
implement
express
sea
joint
effort
posit
current
negoti
exchang
plan
agreement
prime
need
region
polici
summit
dialogu
continu
achiev
abe
held
futur
mutual
confer
situat

remark
import
attack
talk
support
polit
base
area
act
nato
quot
statement
way
regard
make
muslim
air
region
secur
intern
time
spokesman
republ
threat
rouhani
new
agreement
plan
pyongyang
need
meet
persian
relat
arab
end
enemi
terrorist
group
armi
islam
oper
conflict
ministri
target
clear
terror
missil
taliban
leader
council
recent
possibl
territori
provinc
nuclear
offici
issu
dprk
includ
sanction
follow
diplomat
carri
say
korean
posit
action
use
held
militari
arm
polici
continu
power
launch
strike
weapon
unit
usa
deal
joint
crisi
concern
peac
situat
war
regim
fight
respons
defenc
border
troop
accord
secur_council
command
effort
resolut
islam_state
arm_forc

control
offic
coalit
terrorist
offensreturn
presenc
civilian
region
oper
group
camp
war
train
protect
carri
conduct
lavrov
militari_oper
kurdish
weapon
help
secur
includ
provid
conflict
accord
led
continu
rebel
readi
fight
defenc
intern
guard
use
intellig
personnel
number
soldier
attack
strike
base
near
end
situat
kurdistan
launch
militari
arm
refuge
kurd
told
deploy
ground
mission
special
secur_forc
organ
aid
posit
nato
cross
troop
captur
support
battl
statement
hurriyet
zone
command
arm_forc
ong
mosul
air
gen
liber
staff
militia
milit
target
say
plan
armi
border
radio
kill
unit
combat
chief
peac
fighter
area
territori
northern
town
eastern
southern
humanitarian

Aggregate Topic Shares in 4.4 Million Texts

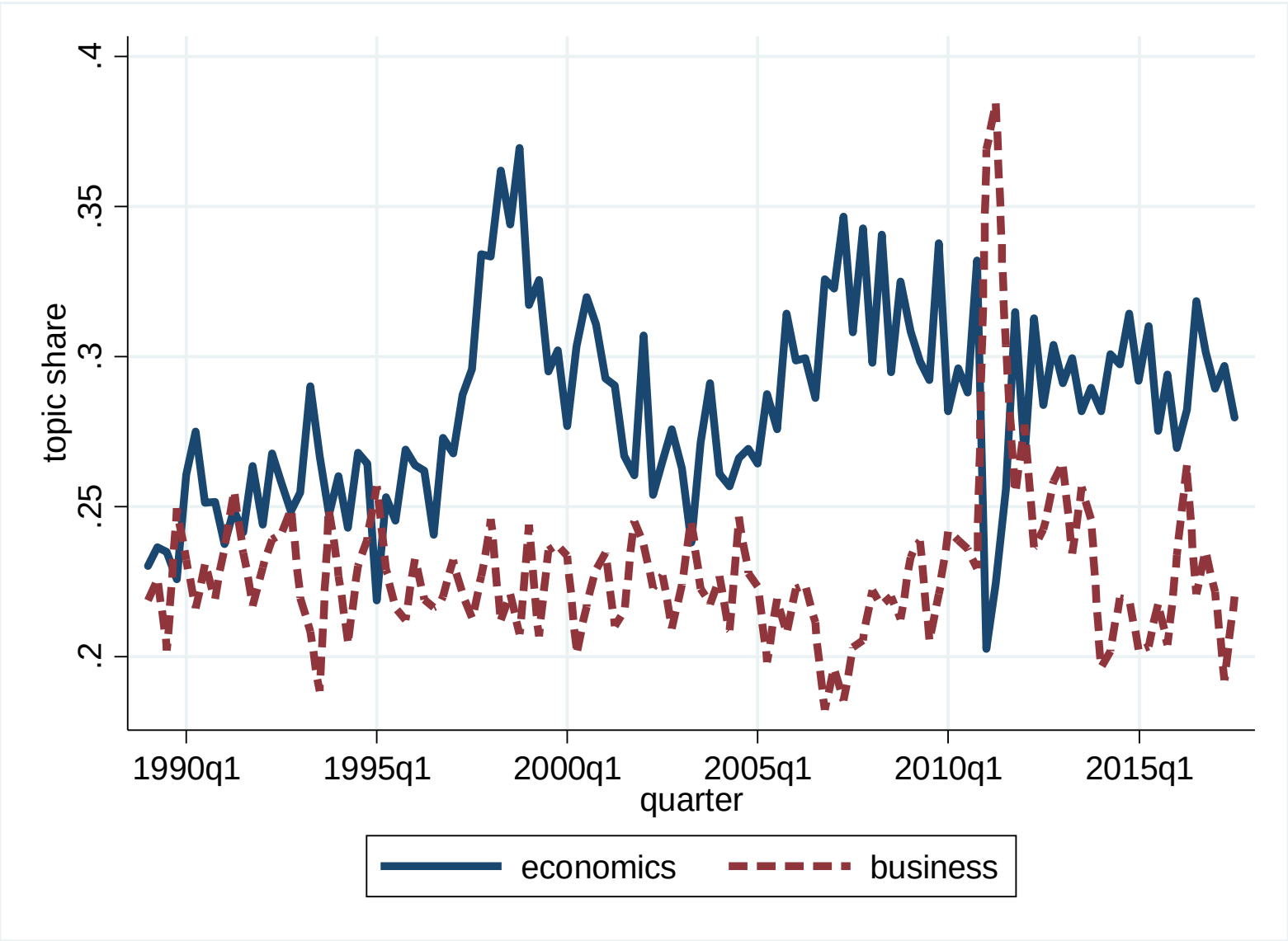
- composition of each article m in terms of the K topics, $\boldsymbol{\eta}_m$
- group of articles written in country i and time t, M_{it}
- *topic shares* in country i in period t is

$$\boldsymbol{\theta}_{it} = \left(\sum_{m \in M_{it}} \boldsymbol{\eta}_m N_m + \alpha\right) / \left(\sum_{m \in M_{it}} N_m + K\alpha\right) \tag{1}$$

  where $\sum_{m \in M_{it}} N_m$ is simply the total number of articles
- α enters here as the strength of the prior
- We estimate topics for each sample $T \in \{2000Q1, 2000Q2, ...2016Q4\}$ separately
- Use quarterly and yearly aggregates $\boldsymbol{\theta}_{it}$ for forecasting.

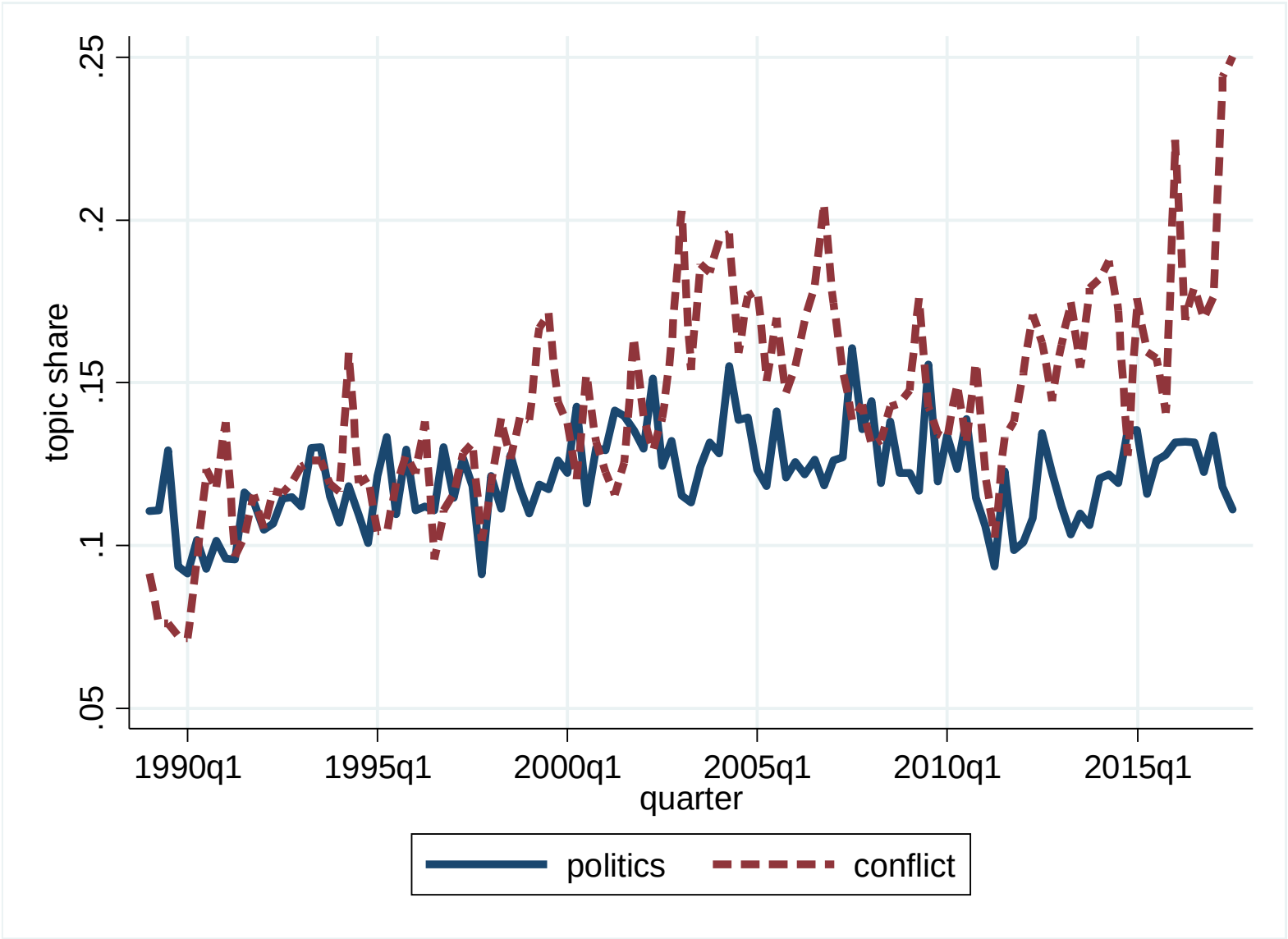
Example of Topic Share Aggregates

"economics and trade politics" and "business and energy" in Japan



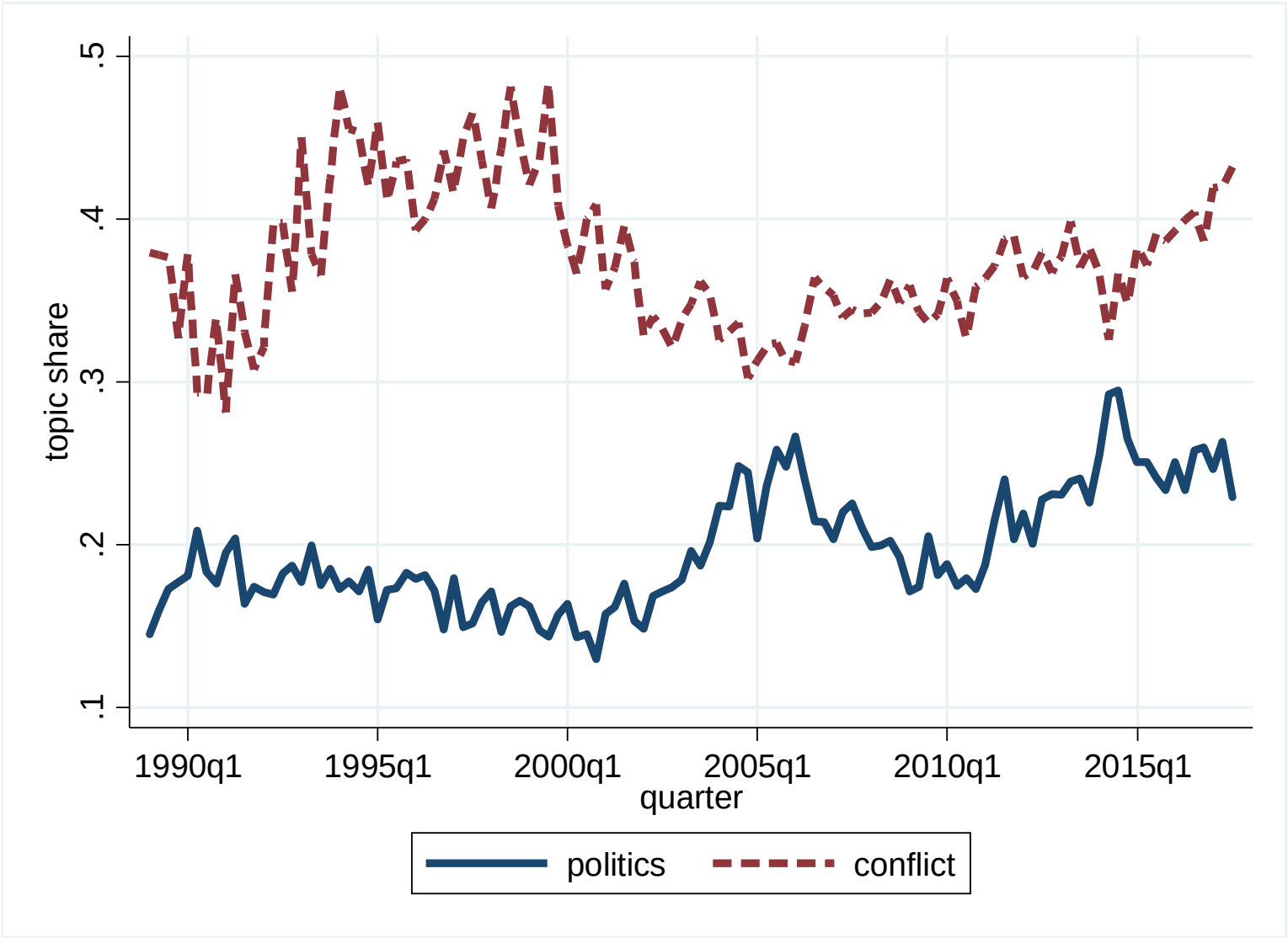
Example of Topic Share Aggregates

"crime and politics" and "conflict/security" in Japan



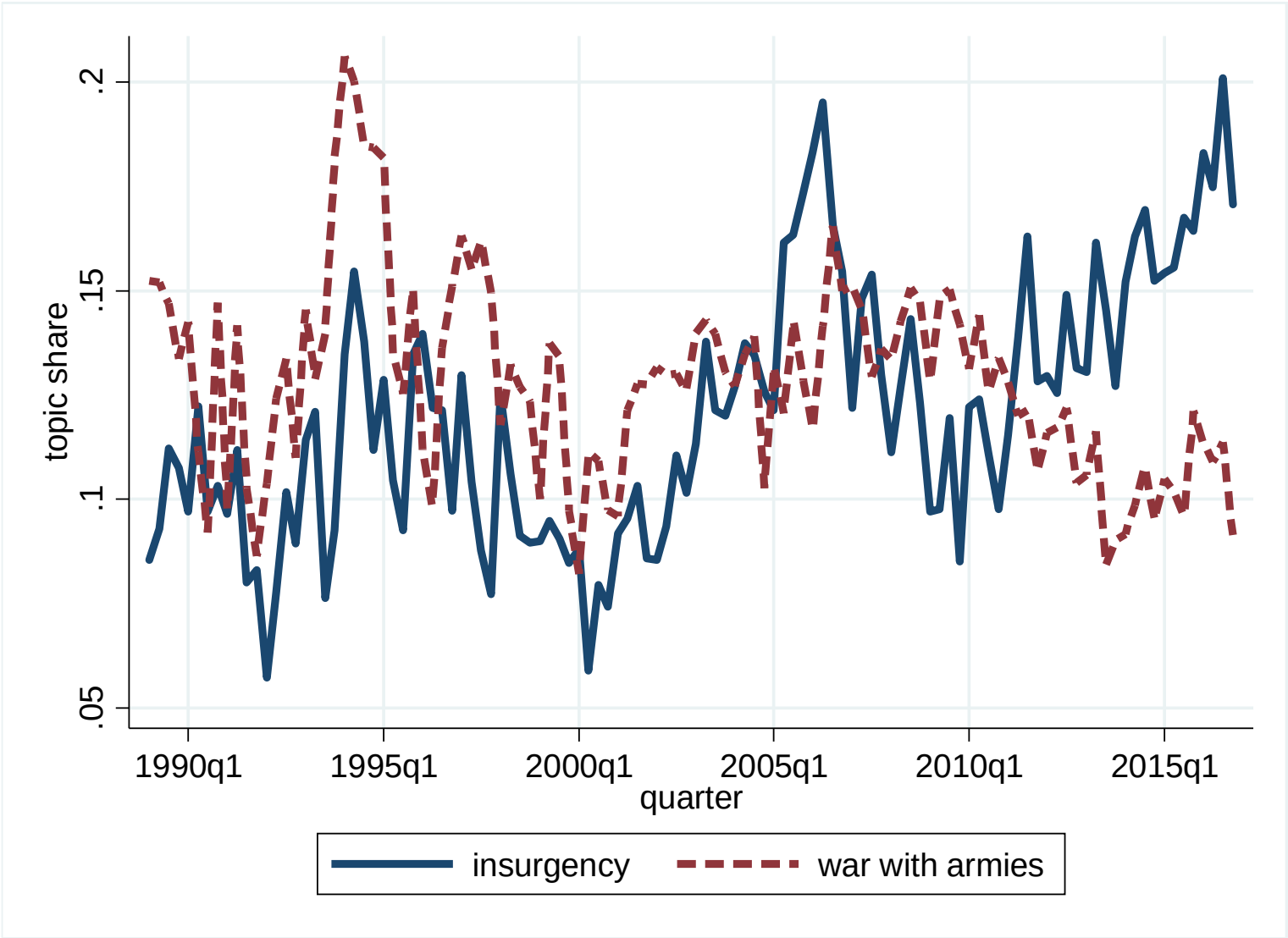
Example of Topic Share Aggregates

"crime and politics" and "conflict/security" in Afghanistan



Example of Topic Share Aggregates

both conflict topics ($K = 15$) in Afghanistan



Forecast Procedure

- Train with data available in T

$$y_{it+1} = F(\mathbf{h}_{it}, \boldsymbol{\theta}_{it}) \tag{2}$$

  where y_{it+1} is the onset of armed conflict in the quarter $t+1$.
- $\mathbf{h}_{it}$: set of history dummies capturing post-conflict dynamics
- $\mathbf{h}_{it}$ also includes dummy for low-level violence and conflict in neighboring countries
- $\boldsymbol{\theta}_{it}$: share of topics
- Calculate predicted values $\hat{y}_{iT+1}$
- Do this for all years $T = 2000Q1 - 2016Q4$.
- To generate $F(.)$ we feed random forests, neural network and logit regression into an ensemble

Benchmark Model

- Use the conflict history dummies as a benchmark model.

$$y_{it+1} = F(\mathbf{h}_{it}) \tag{3}$$

- How much does text add beyond this benchmark?
- Prior: text will add more in the hard problem

How is F derived?

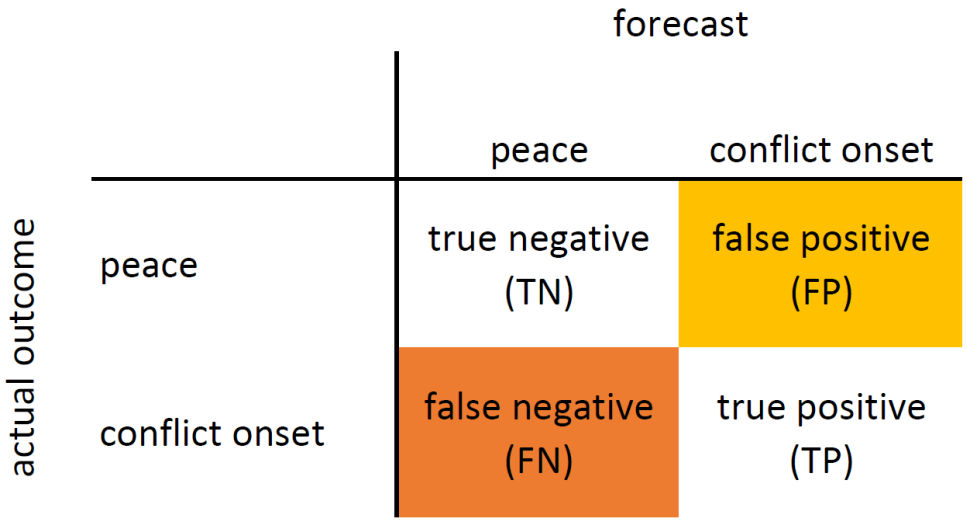
- We use machine learning to derive $F(\cdot)$
- This is not because it improves the fit within-sample.
- It's because we want to use the $F(\cdot)$ to produce fitted values for $\hat{y}_{iT+1}$.

How is F derived?

- 1) we estimate a logit model
- 2) we estimate a random forest: depth 3 and 175 estimators.
- 3) we estimate a neural network: 3 layers and the number of neurons are 2, 5 and 7.
- Weight between 1), 2) and 3) is built with ensemble through soft voting.

Results: Trade-offs in Forecasting

- Need to decide on cut-off c to evaluate model:
- $\hat{y}_{iT+1} > c \longrightarrow$ forecast conflict
- Two mistakes we can make:



- high cutoff c implies more false negatives
- low cutoff c implies more false positives

Results: Standard View

- ROC curves as a way to illustrate the trade-off.
- On the y-axis report the true positive rate (TPR)

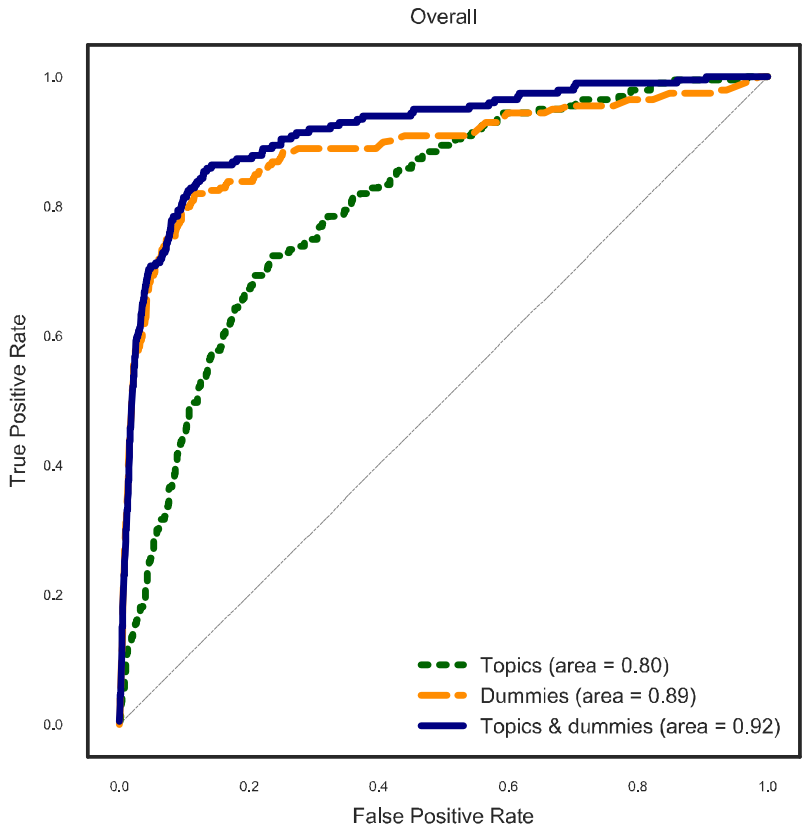
$$TPR_c = \frac{TP_c}{FN_c + TP_c}$$

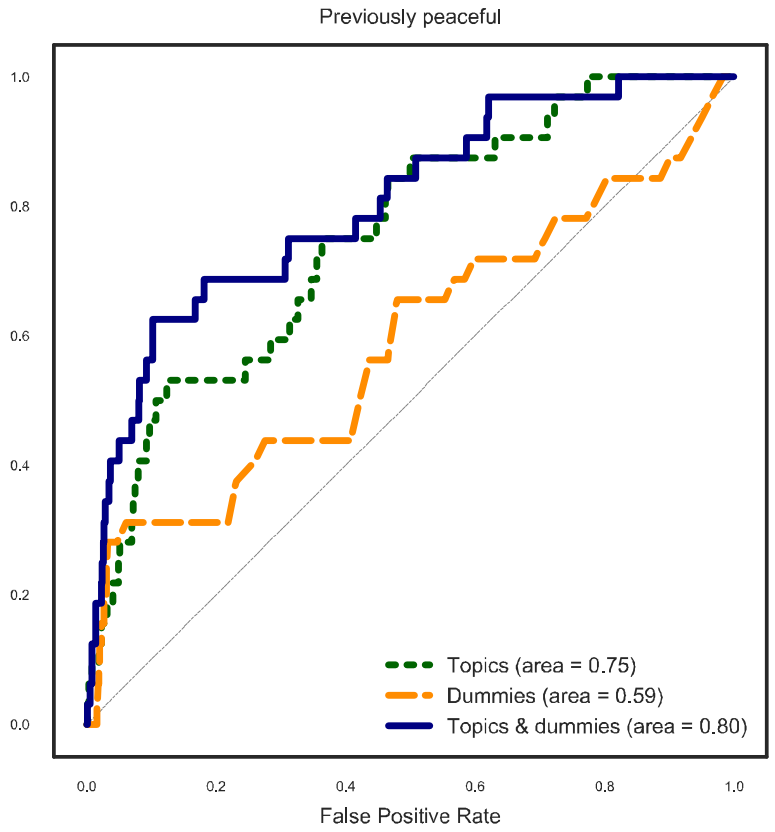
- On the x-axis report the false positive rate (FPR)

$$FPR_c = \frac{FP_c}{FP_c + TN_c}$$

ROC curve: 50+ onsets

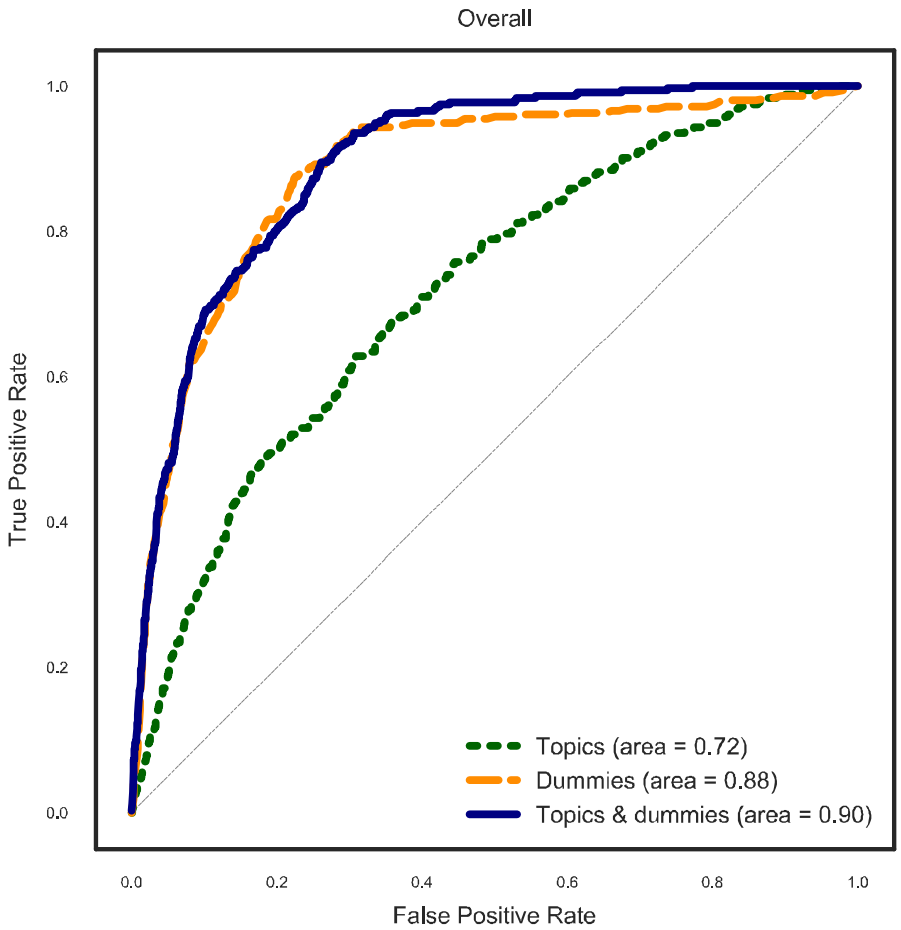
Topics help solve the hard problem

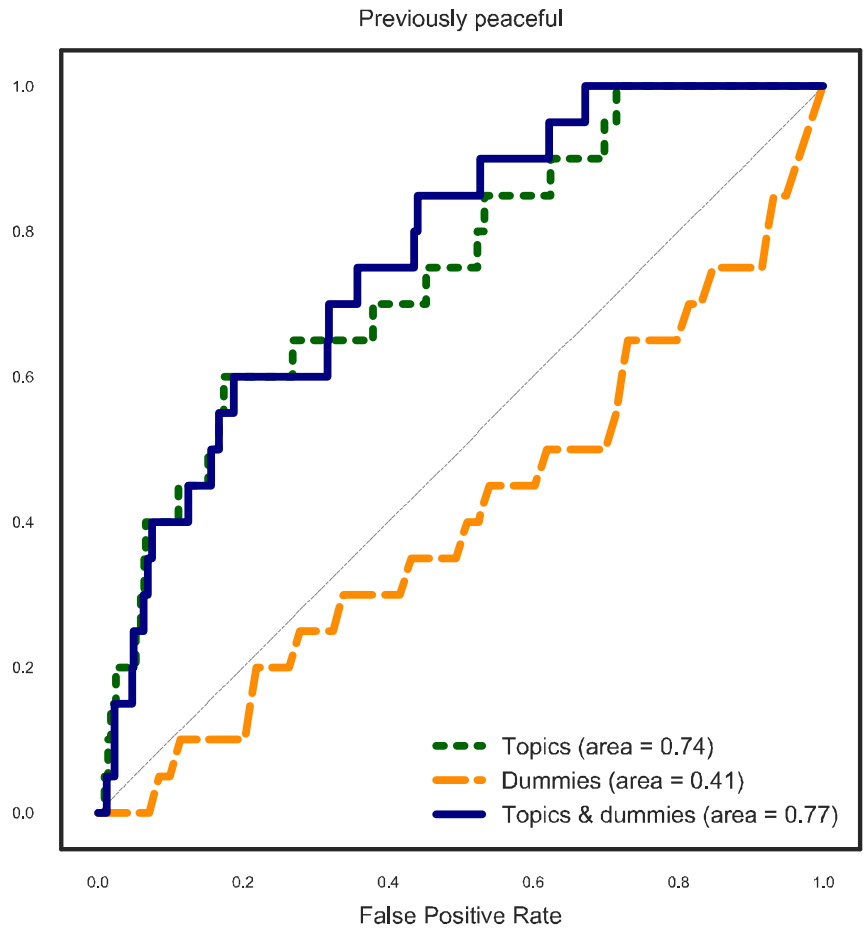




ROC curve: for any violence

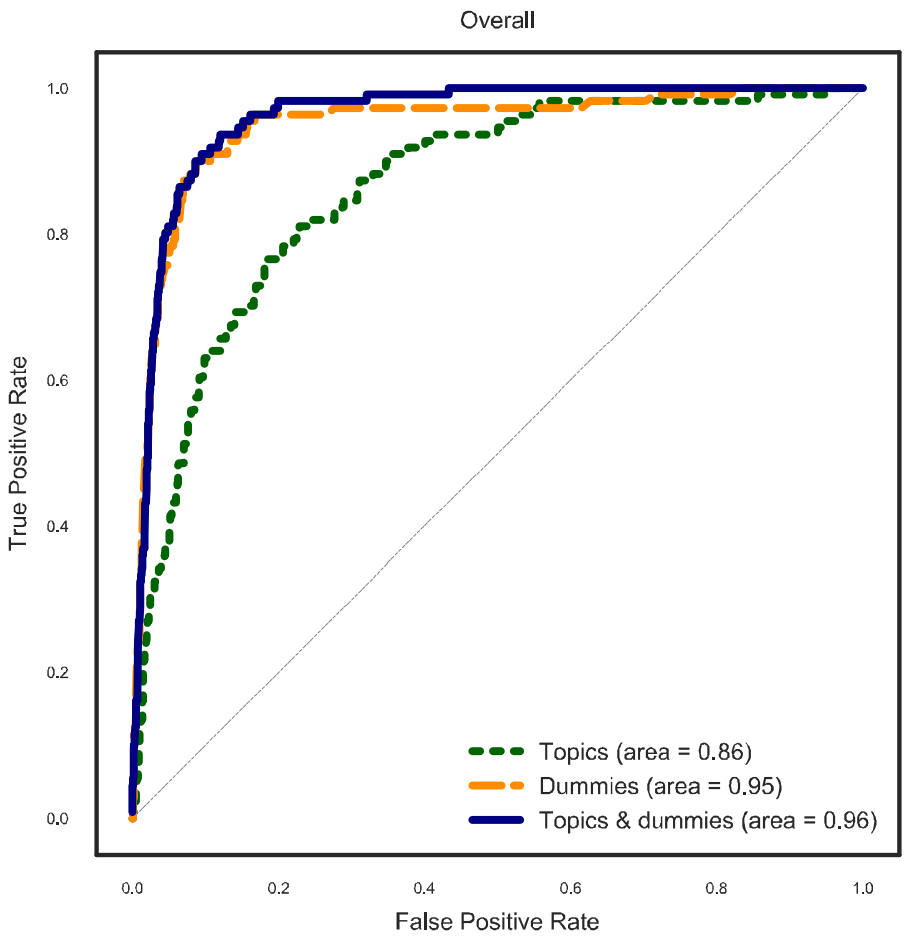
Predicting onset of any violence is harder...

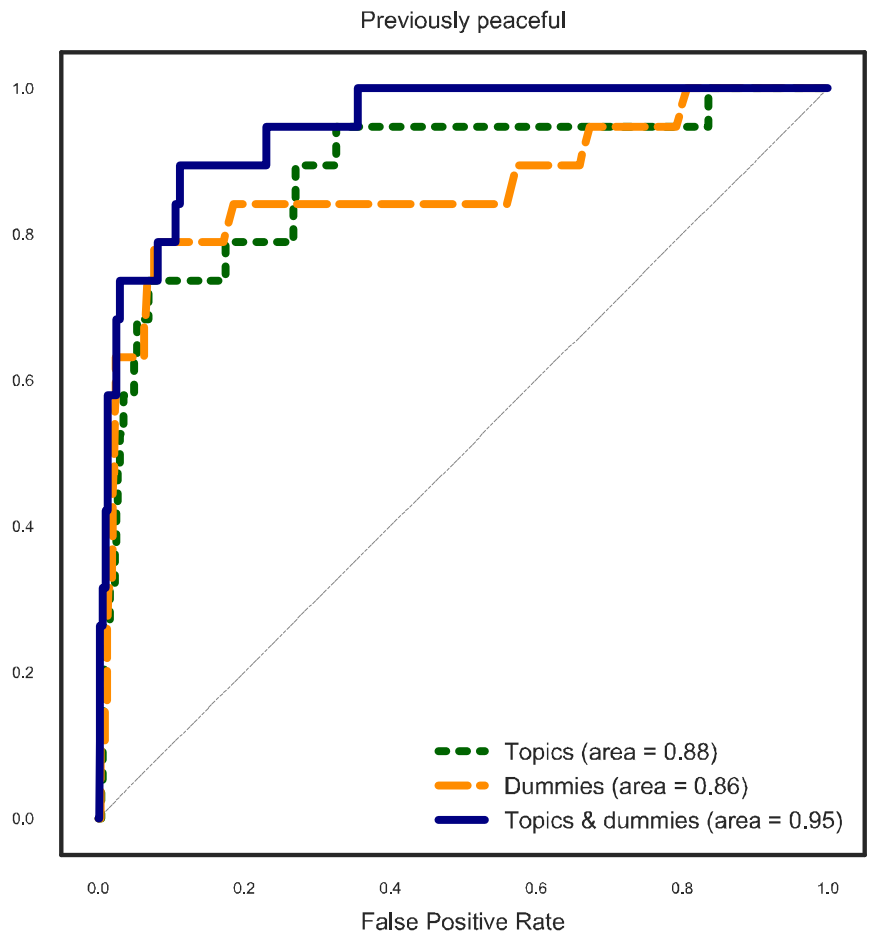




ROC curve: for 500+ onset

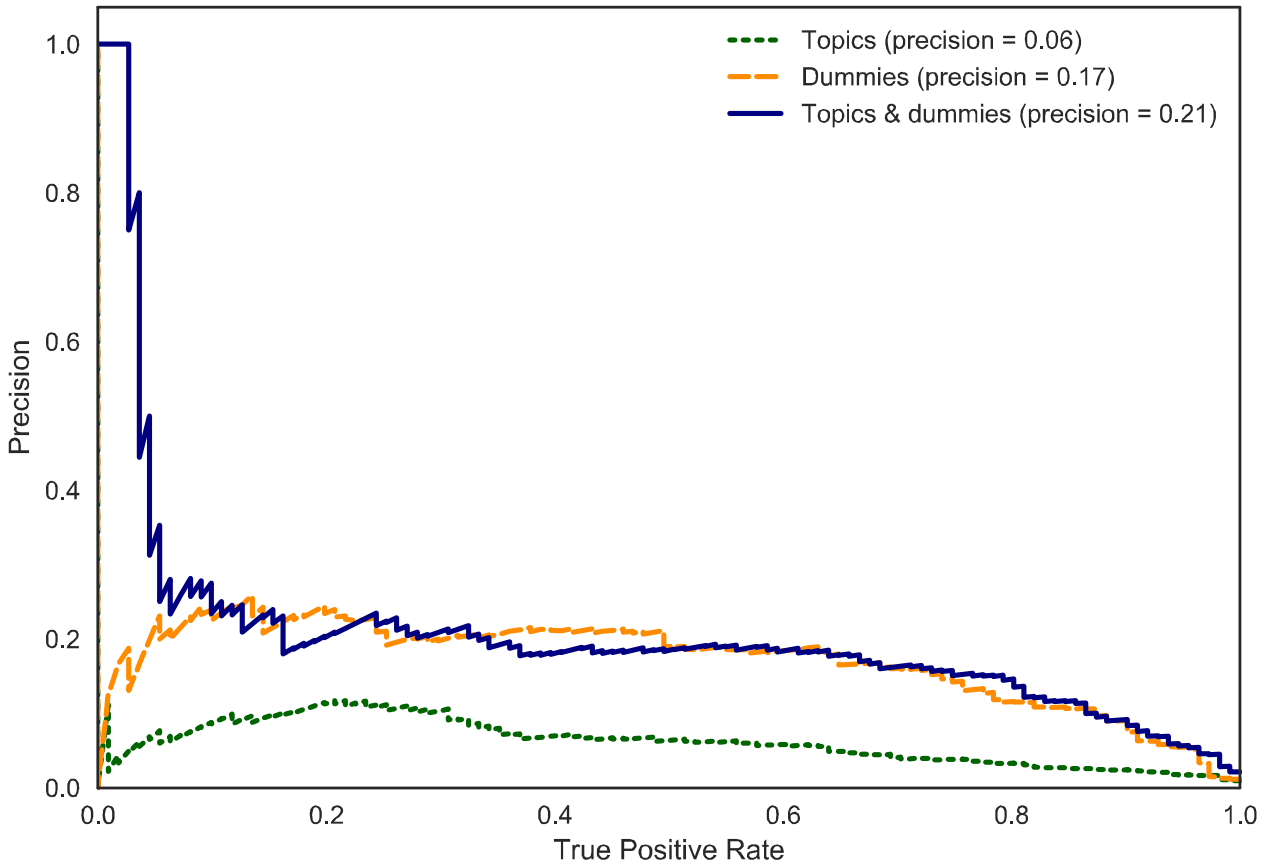
Predicting 500+ escalations is easlier.





Precision vs. True Positive Rate: 500+

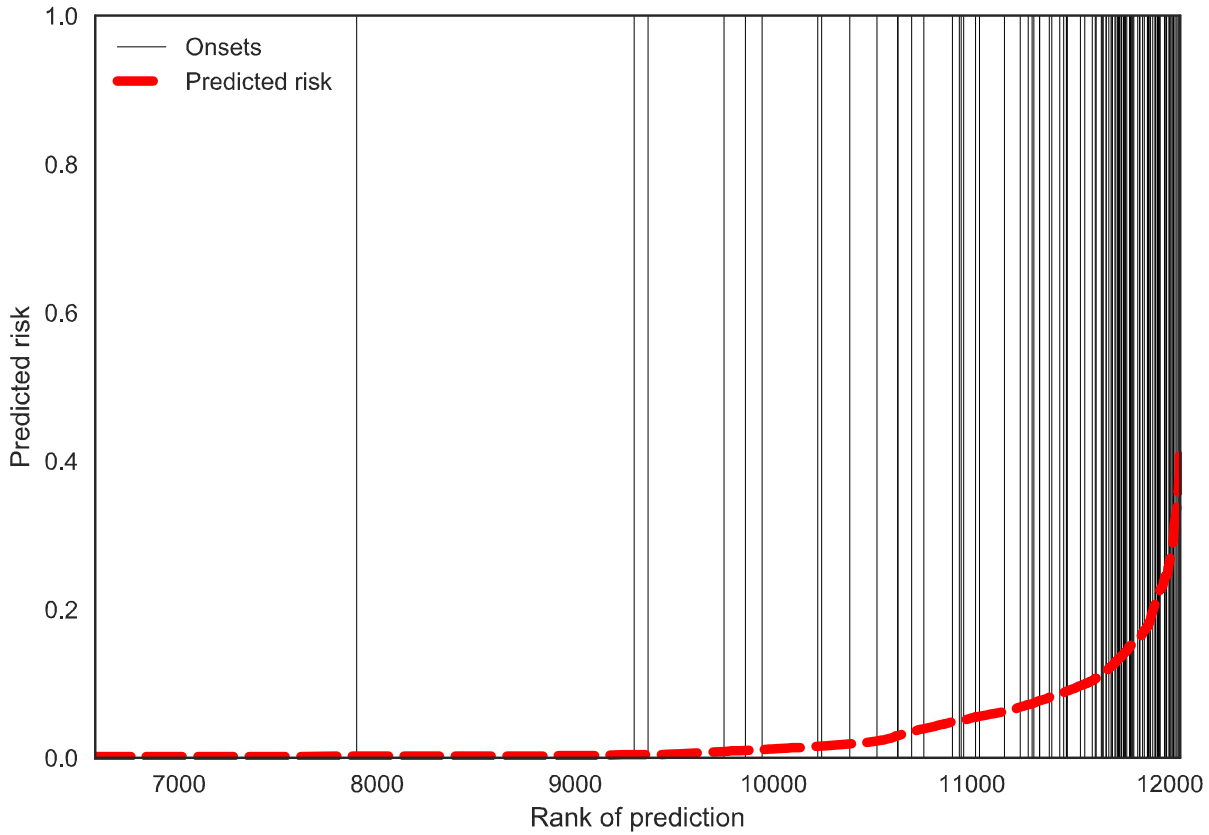
- precision: $P_c = \frac{TP_c}{TP_c + FP_c}$



- 20%: of 5 positives 1 will be a true one.
- In this literature this is really good.

Separation Plot: 500+

- Separation Plot



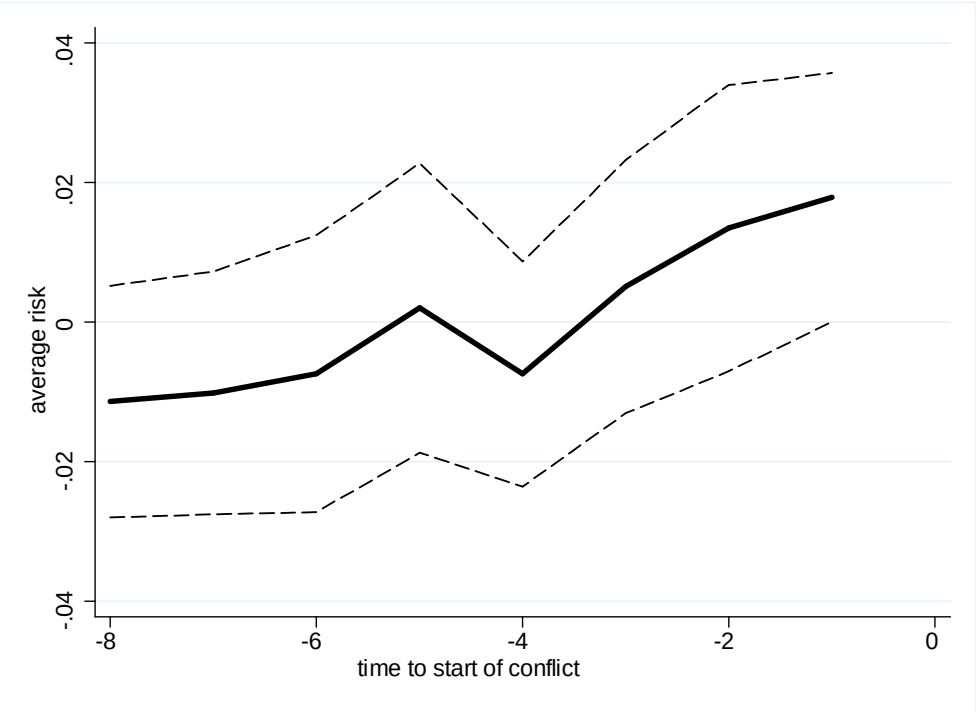
Why does it work?

- machine learning like random forest is using different samples in the training
- methods are non-linear by design (example tree)

insert tree

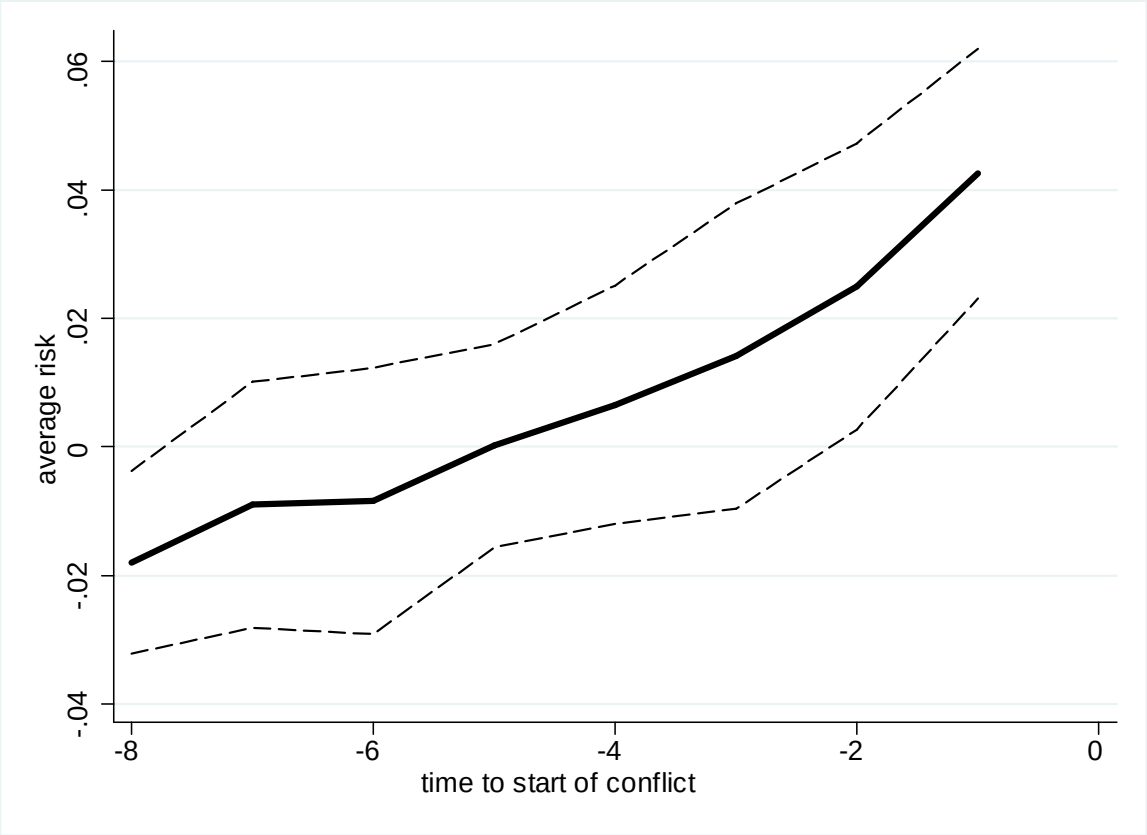
Why does it work?

- text contains some time variation
- plot predicted risk changes before onset (any violence)



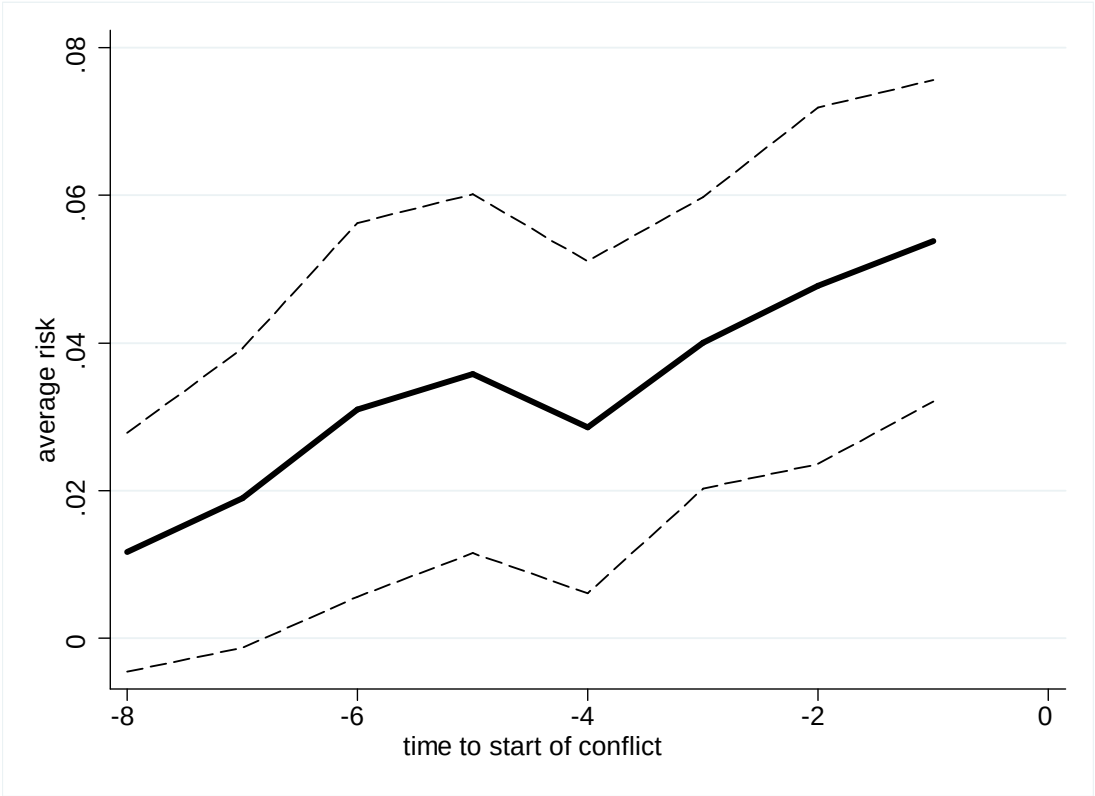
Why does it work?

- plot predicted risk changes before onset (50+ conflict)



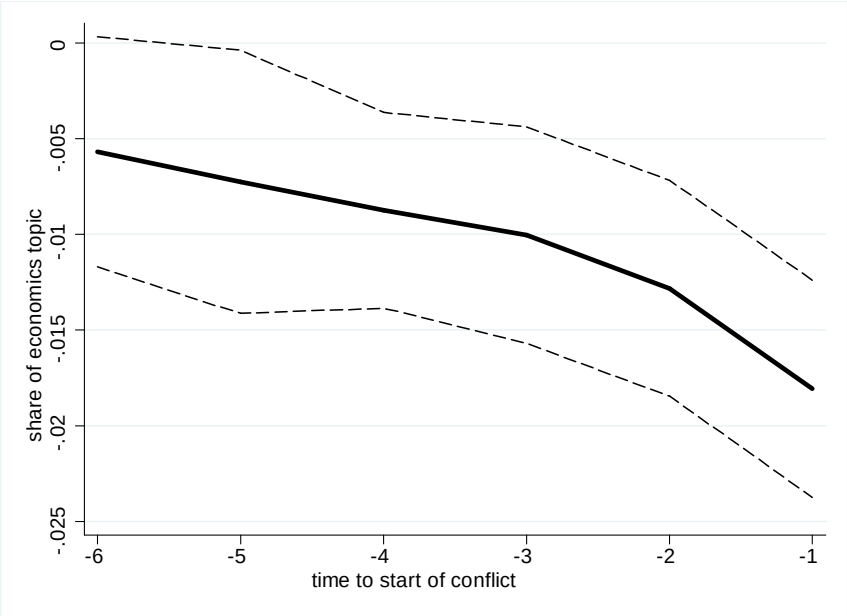
Why does it work?

- plot predicted risk changes before onset (500+ conflict)



Why does it work?

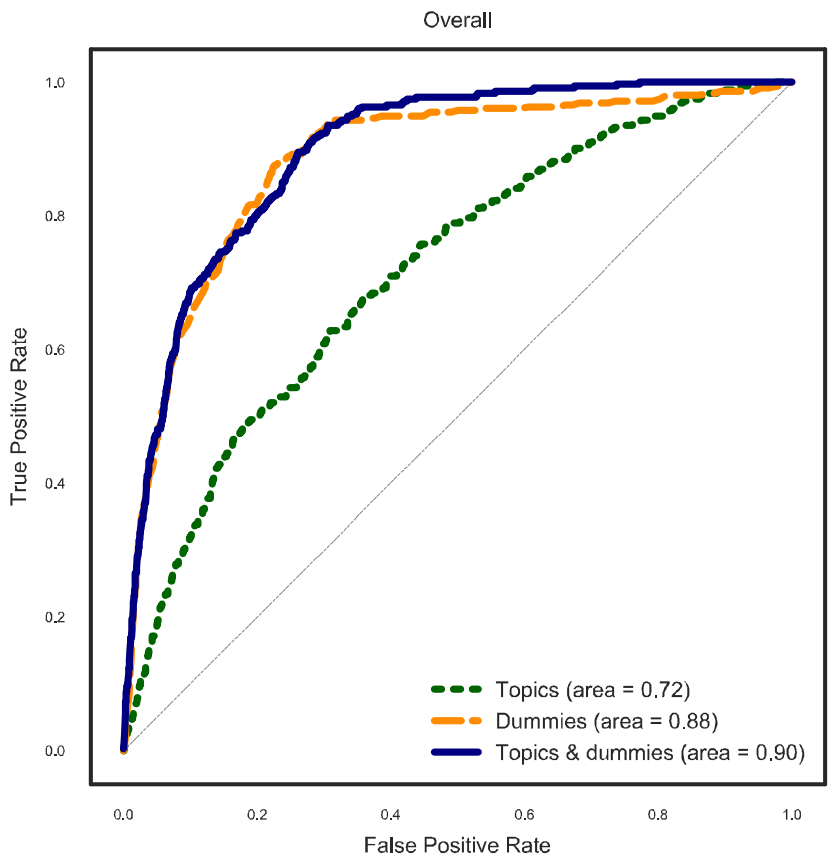
- key is behavior of topics before onset: conflict topics go up before
- but "economics and crime", for example,

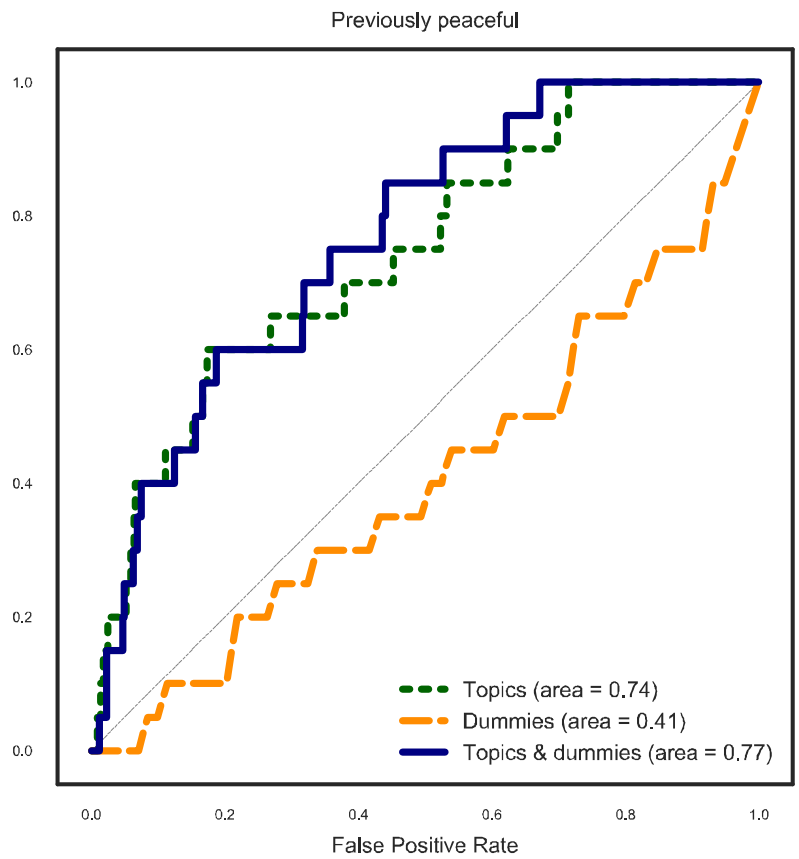


- This is even true when controlling for conflict share.

Back to ROC: 15 Topics

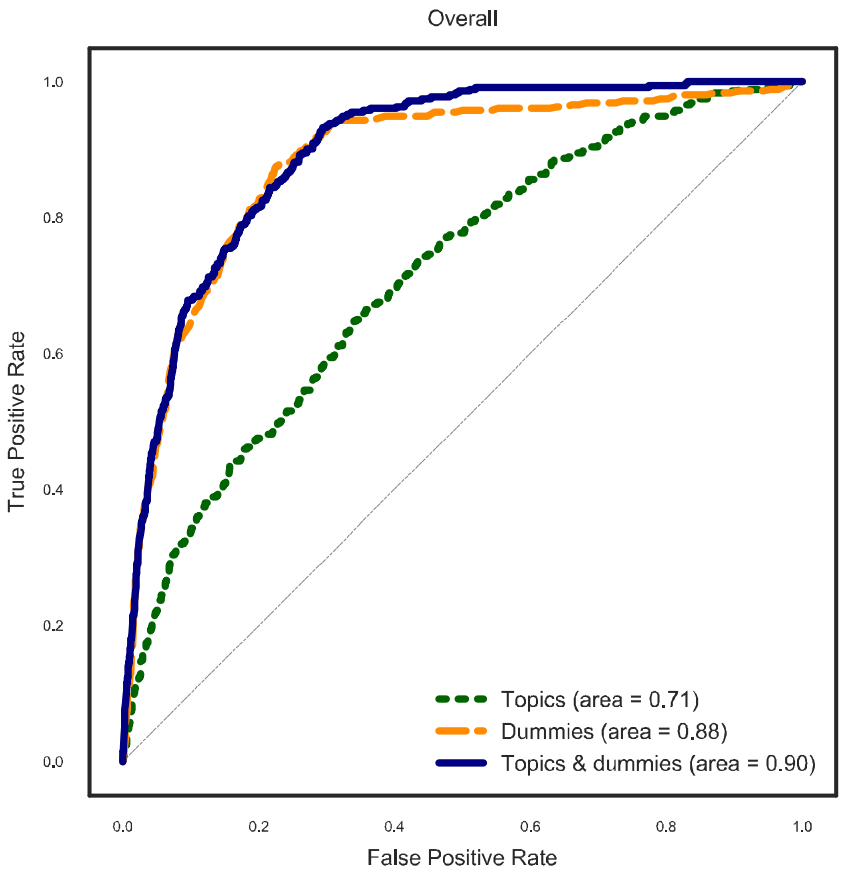
Onset of 1+ conflict

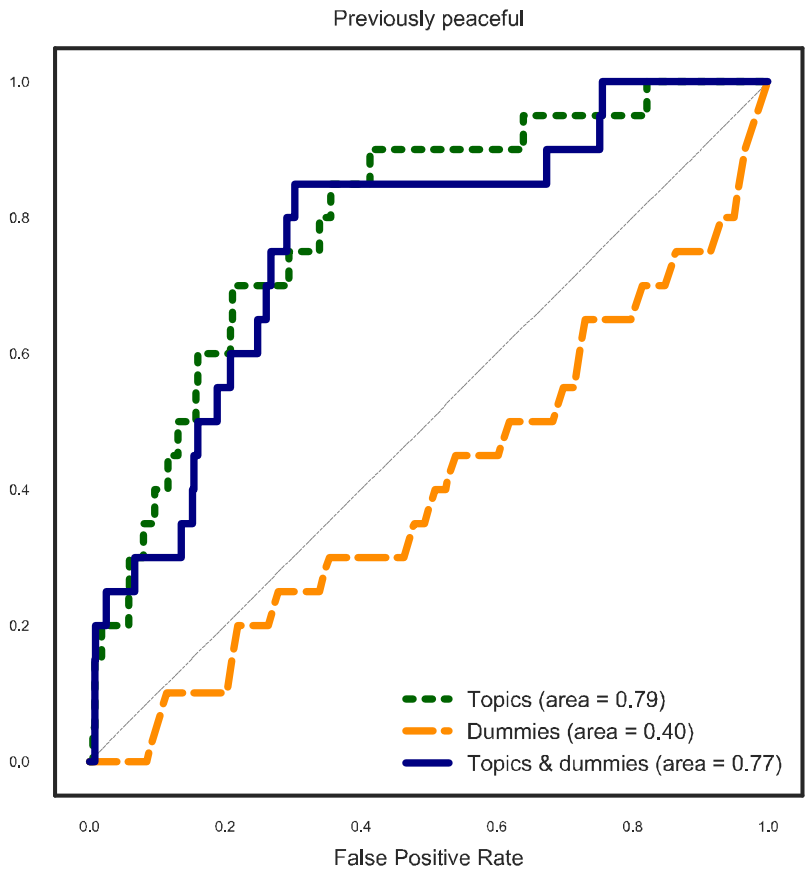




What Happens With Less Topics? (10)

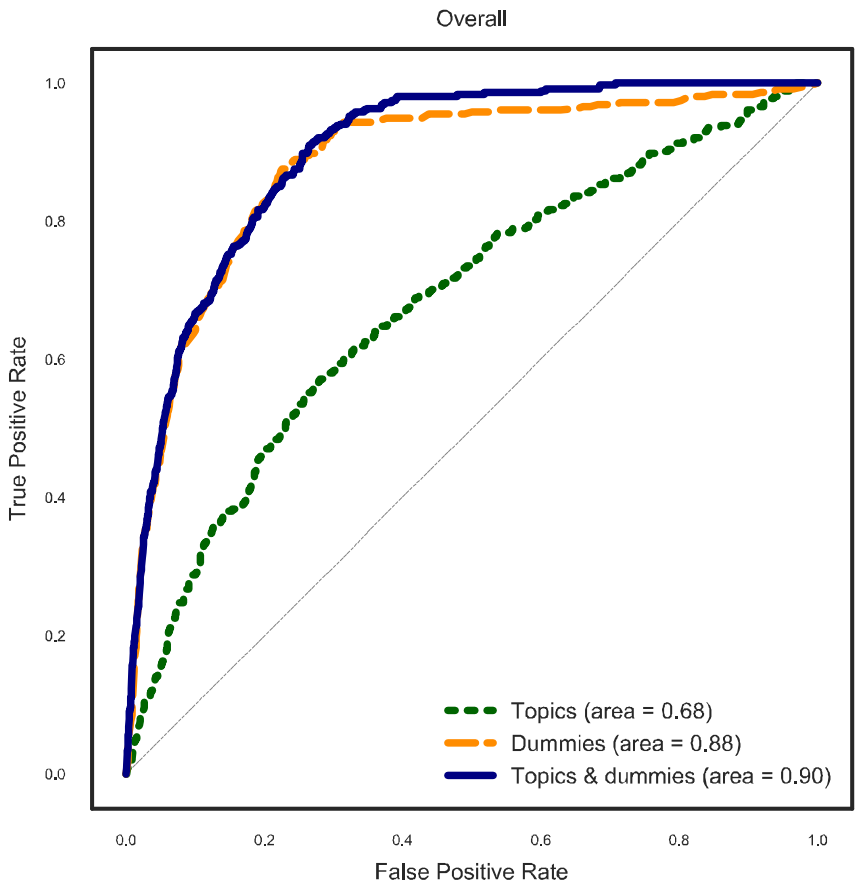
Onset of 1+ conflict

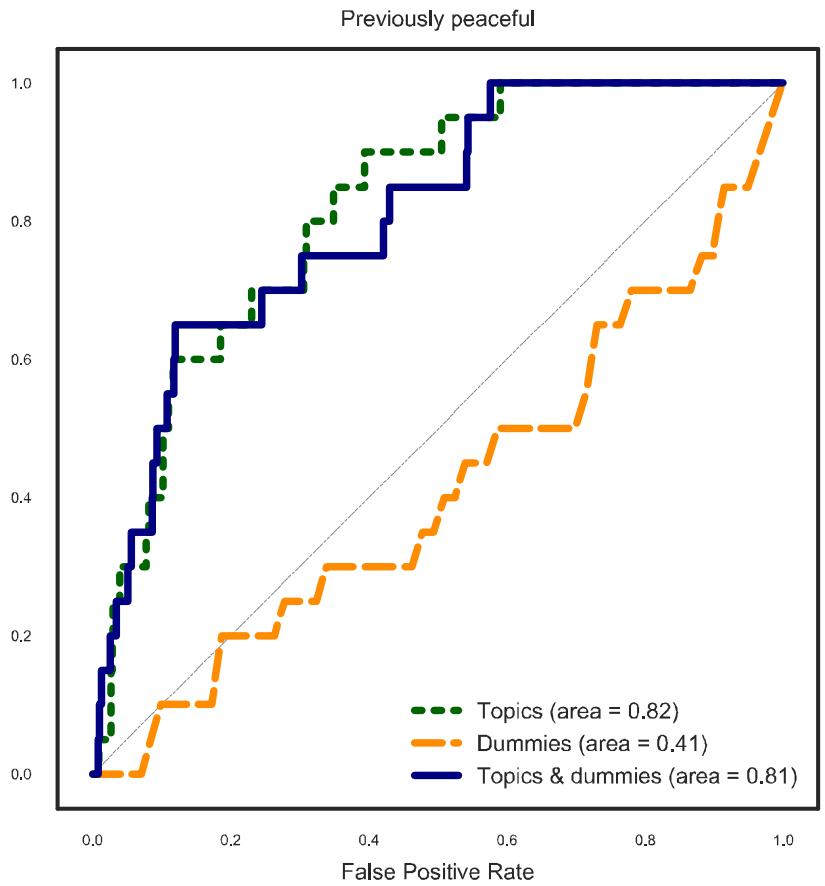




What Happens With Less Topics (5)?

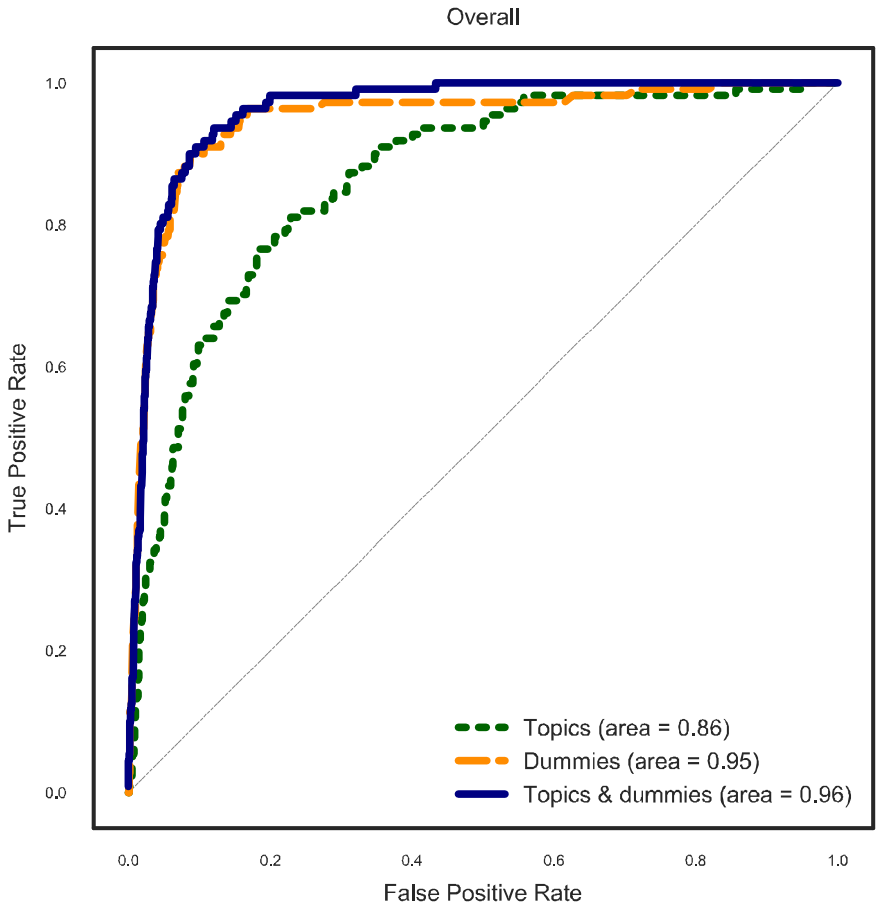
Onset of 1+ conflict

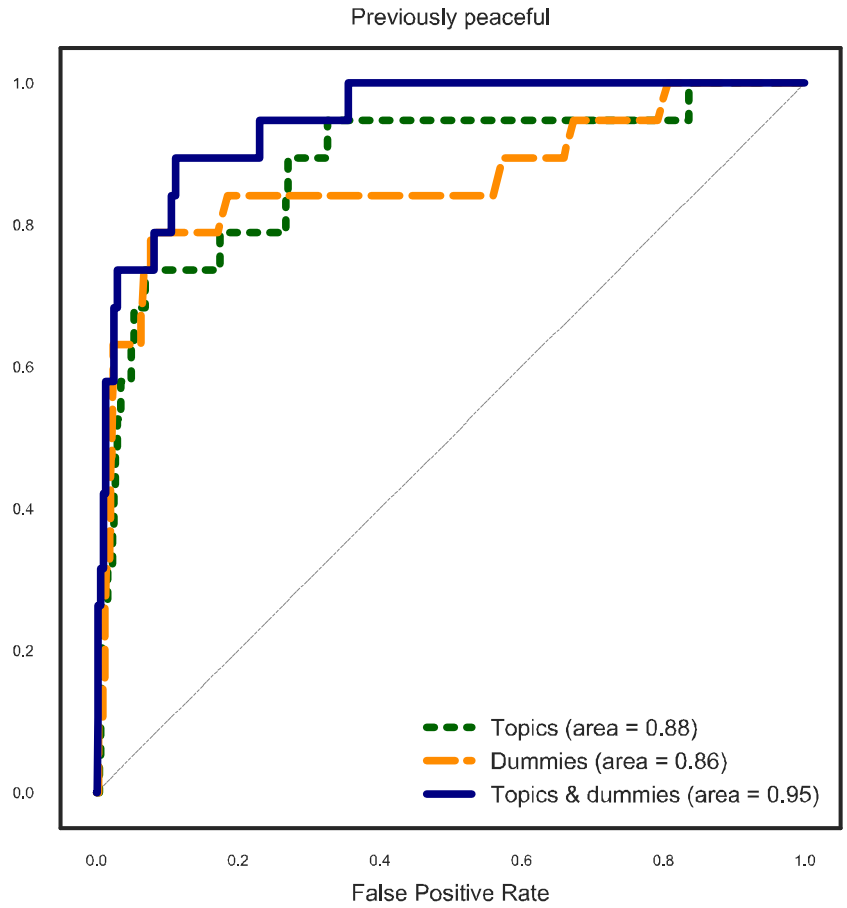




More to learn, more topics are better

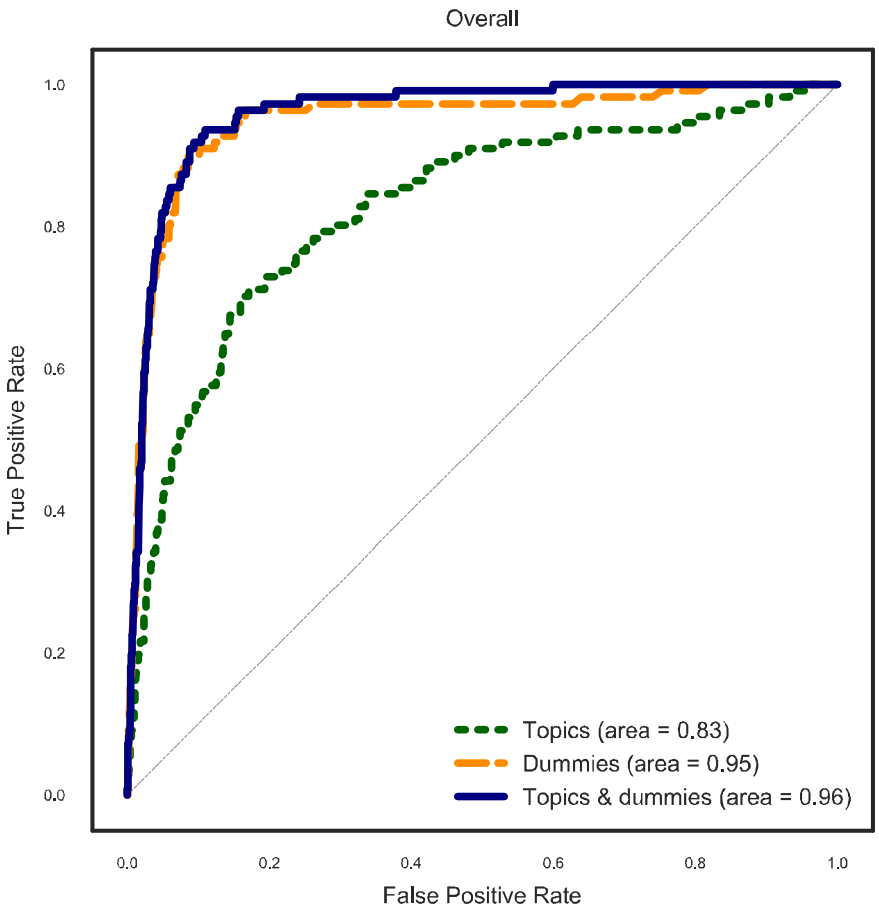
Onset of 500+ conflict, 15 topics

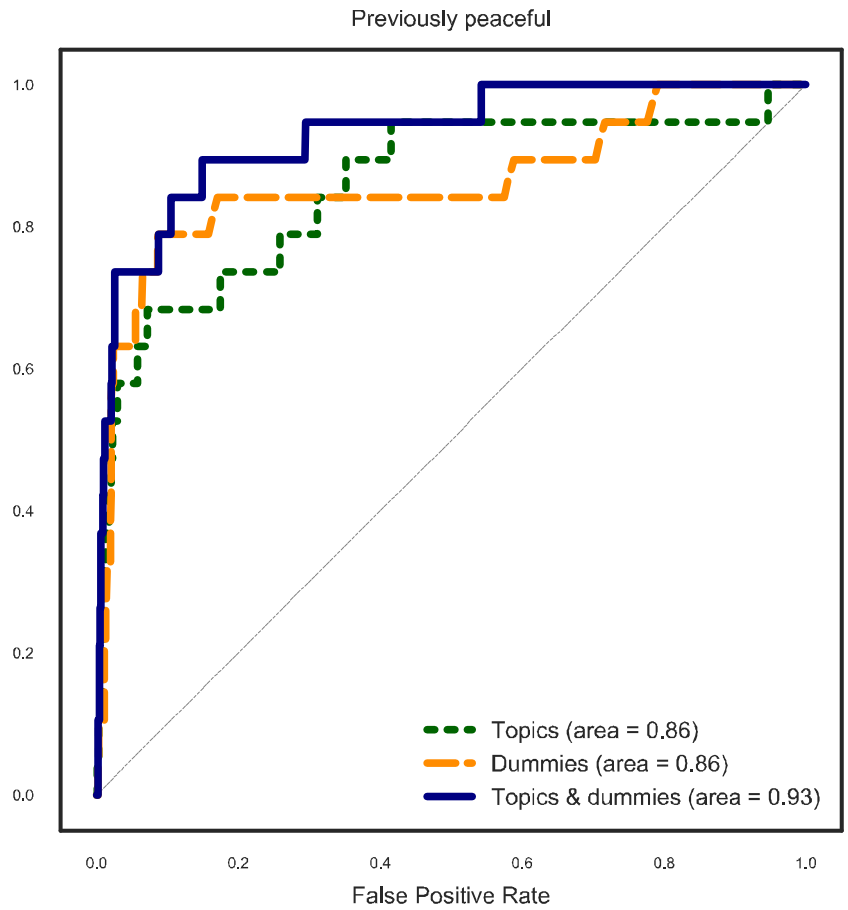




More to learn, more topics are better

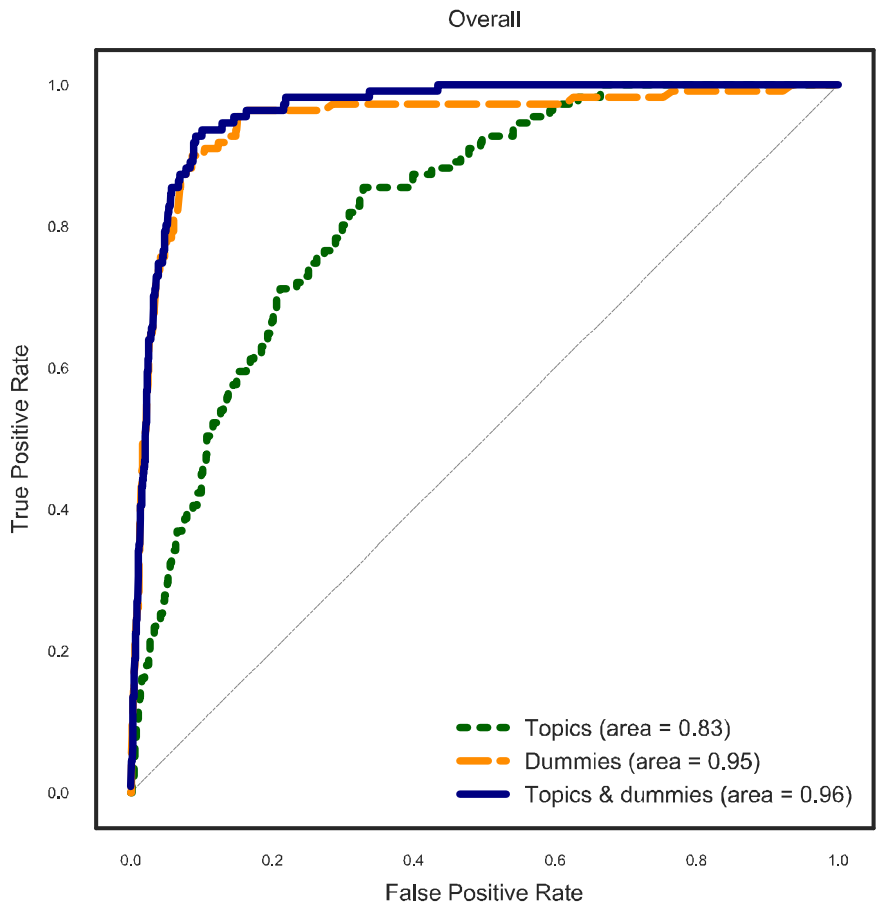
Onset of 500+ conflict, 10 topics

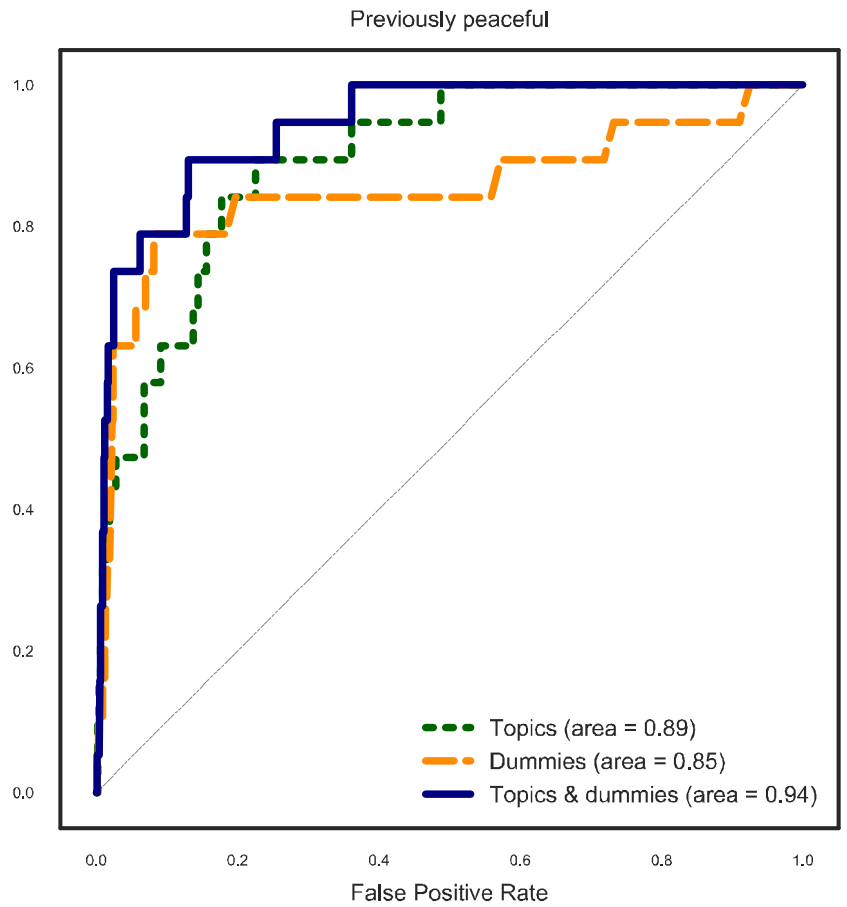




This is not general, however.

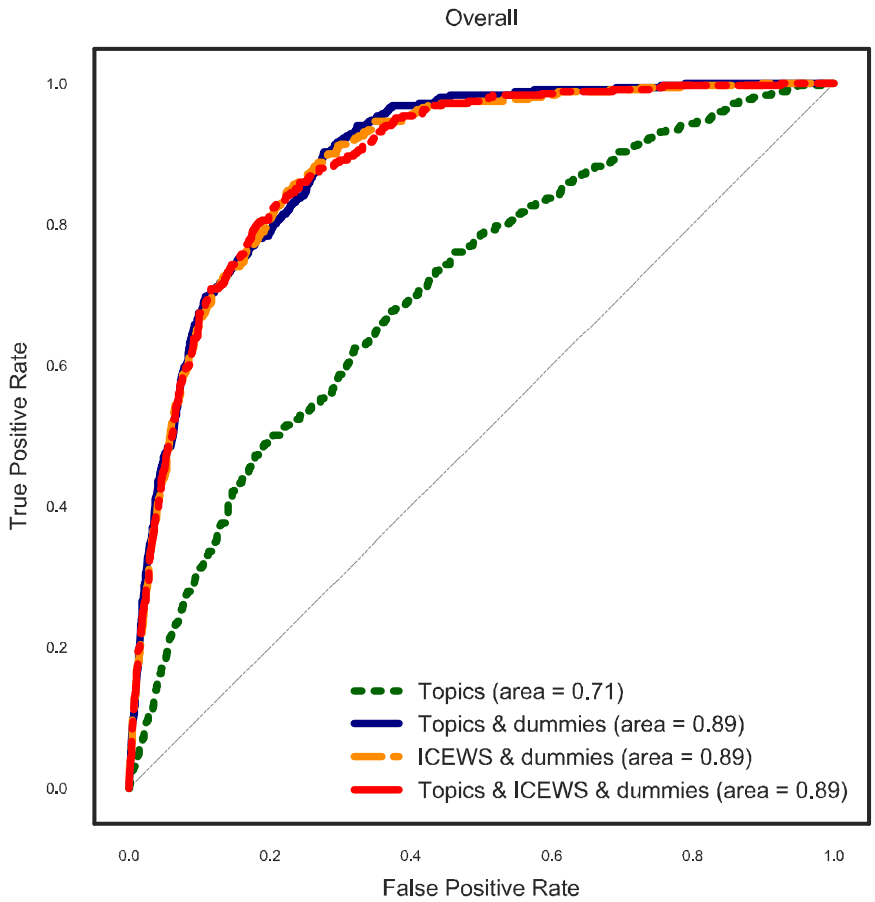
Onset of 500+ conflict

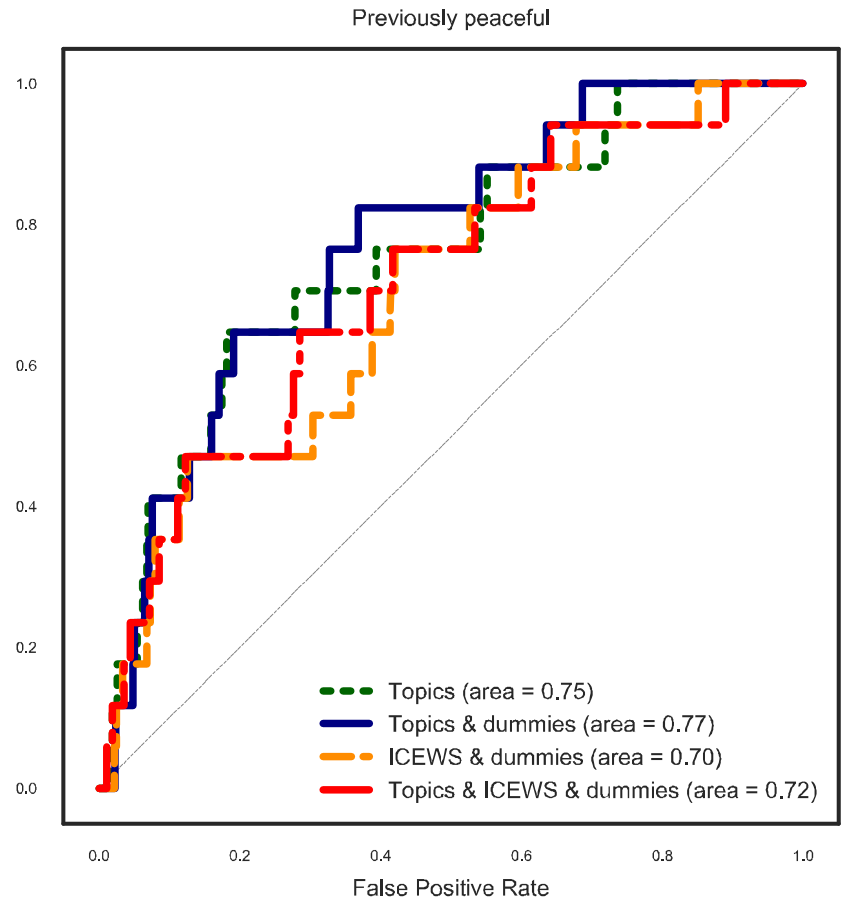




Compare to ICEWS

Event database of more than *6 million events*





Discussion

- conflict history is a great forecaster
- without conflict history forecast becomes a *hard problem*
- use topics to summarize massive amounts of newspaper text
- topics provide some forecast without a history
- lessons for use of unsupervised vs. supervised learning