The Language of Manipulation: Propaganda and Computational Solutions

Anjalie Field, anjalief@cs.cmu.edu

Recap from Tuesday

- Some propaganda strategies are overt
  - Demonize the enemy
  - Fake news

- Some strategies are more subtle
  - Obfuscate the source
  - Flood with misinformation

Recap from Tuesday

- What is the role of technology in propaganda?
  - Twitter is a forum for public opinion manipulation
  - Dangers (real or not…) of automated content generation
  - Often perceived negatively
- Can technology have a positive impact?
  - “Fake News” detection and fact checking
  - What about more subtle strategies of propaganda?
    - Automated analysis of strategies
    - A “propaganda classifier”
    - If we can do it for hate speech, can we do it for propaganda?

Overview: Towards Computational Solutions

- Step 1: What types of manipulation strategies do we see in modern era?
  - We need to know what we're looking for!
  - Ground truth data leaks have given us some insight
  - Strategies are often subtle and hard to detect

- Step 2: What can we learn from social science theories of public opinion manipulation?

- Step 3: How can we use this information to automate propaganda detection and analysis?

Public Opinion Manipulation on Chinese Social Media (Step 1a)

How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not engaged argument

- In 2014 email archive was leaked from the **Internet Propaganda Office** of Zhanggong

- Reveal the work of “50c party members”: people who are paid by the Chinese government to post pro-government posts on social media

King, Gary, Jennifer Pan, and Margaret E. Roberts. "How the Chinese government fabricates social media posts for strategic distraction, not engaged argument." *American Political Science Review* 111.3 (2017): 484-501.

Sample Research Questions [King et al. 2017]

- **When** are 50c posts most prevalent?
- What is the **content** of 50c posts?
- What does this reveal about overall government strategies?

- Additionally:
  - Who are 50c party members?
  - How common are 50c posts?

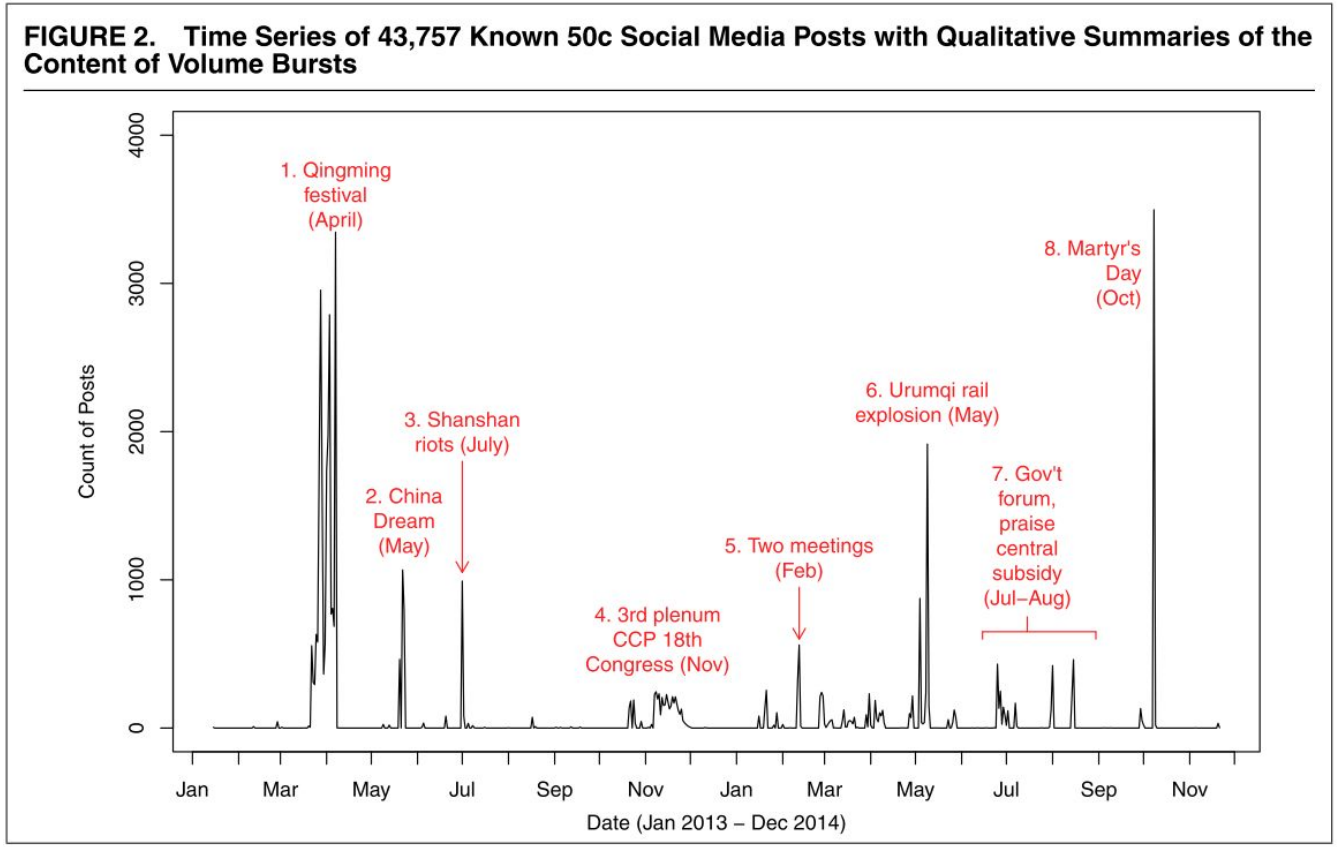
Preparations [King et al. 2017]

- Thorough analysis of journalist, academic, social media perceptions of 50c party members

- Data Processing
  - Messy data, attachments, PDFs

Preliminary Analysis [King et al. 2017]

- Network structure
- Time series analysis: posts occur in bursts around specific events

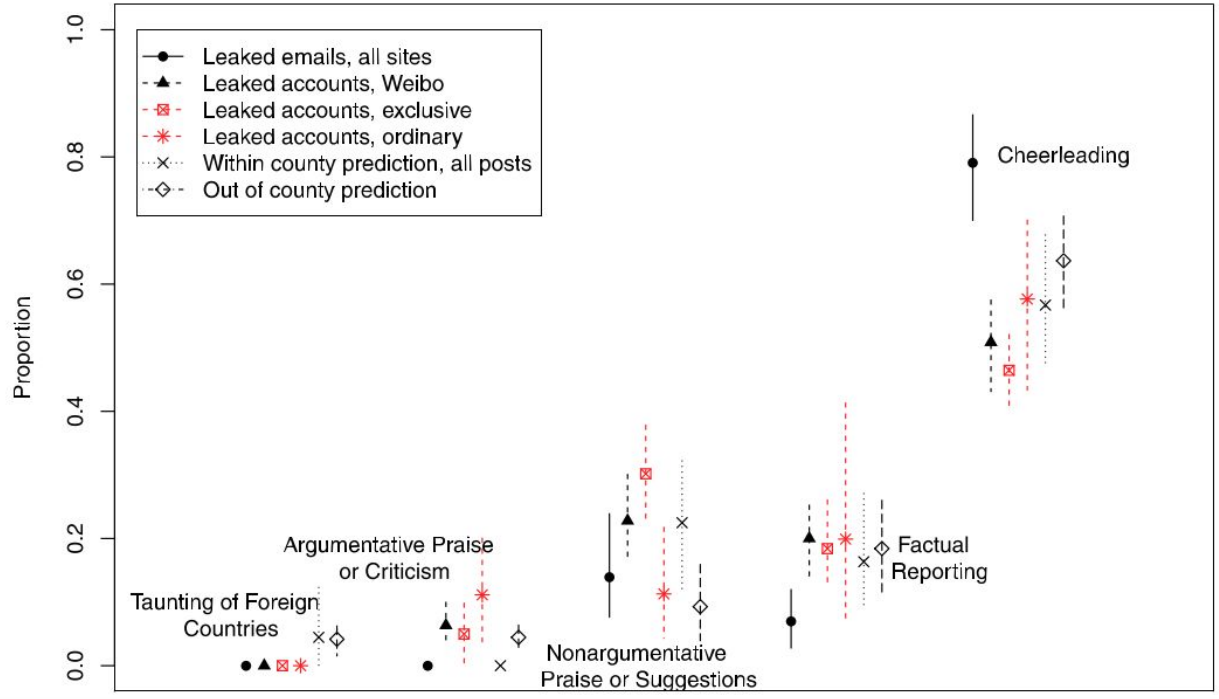


Content Analysis [King et al. 2017]

- Hand-code ~200 samples into content categories
  - Cheerleading, Argumentative, Non-argumentative, Factual Reporting, Taunting Foreign Countries
  - Coding scheme is motivated by literature review
  - Use these annotations to estimate category proportions across full data set
- Expand data set
  - Look for accounts that match properties of leaked accounts
  - Repeat analyses with these accounts
  - *Conduct surveys of suspected 50c party members*

Content Analysis [King et al. 2017]

FIGURE 3. Content of Leaked and Inferred 50c Posts, by substantive category (with details in Appendix A) and analysis (given in the legend)



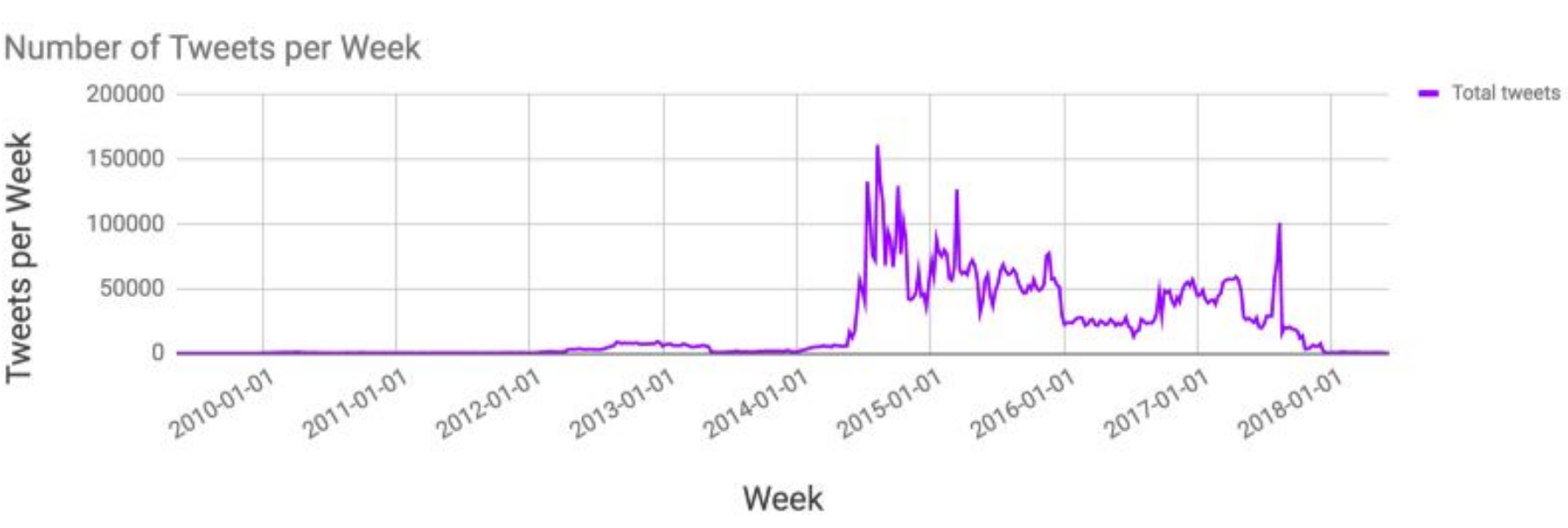
Cheerleading: Patriotism, encouragement and motivation, inspirational quotes and slogans

Public Opinion Manipulation by Russian government on U.S. Social Media (Step 1b)

Twitter recently released troll accounts

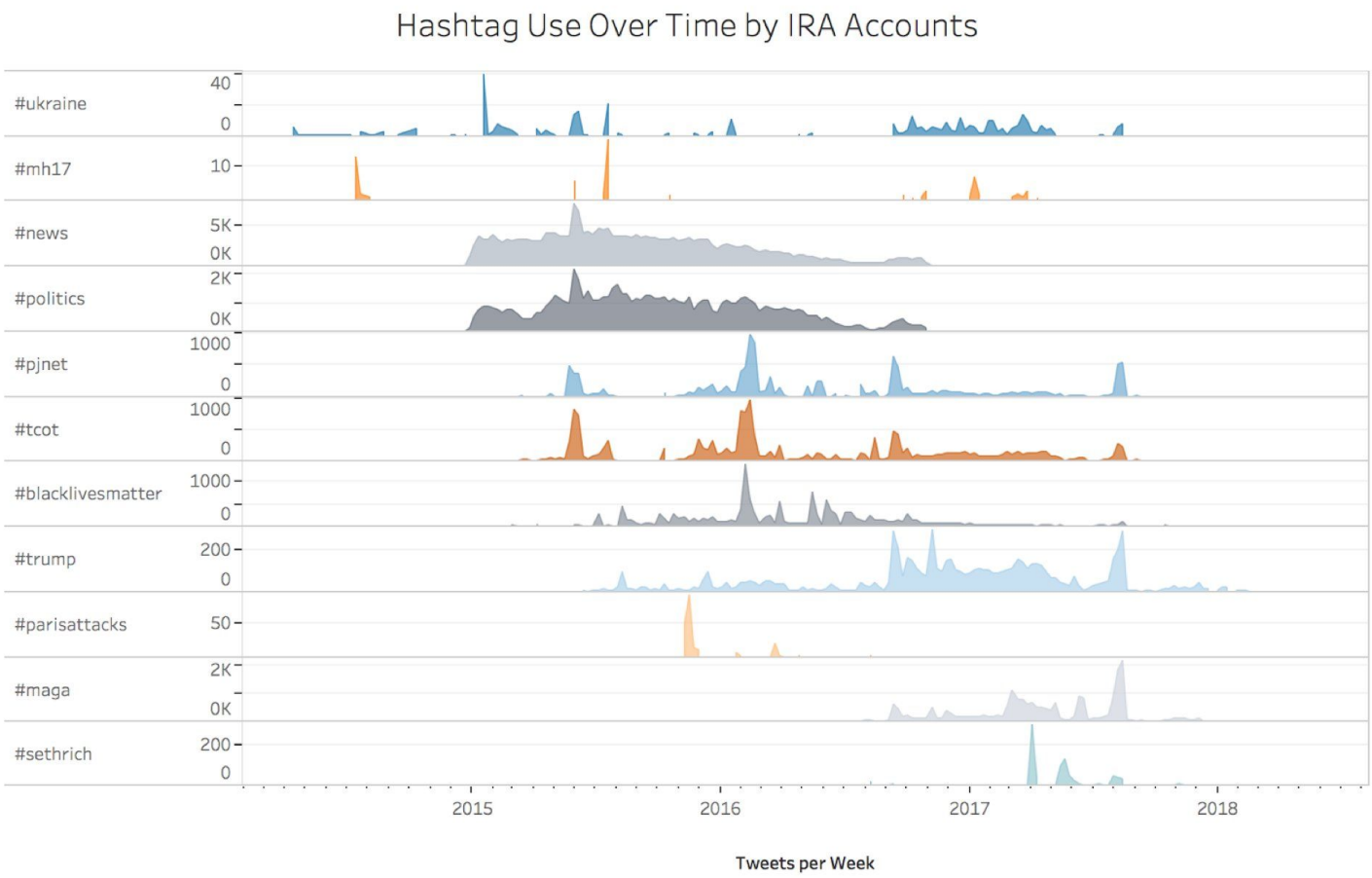
- Information from 3,841 accounts believed to be connected to the Russian Internet Research Agency, and 770 accounts believed to originate in Iran
- 2009 - 2018
- All public, nondeleted Tweets and media (e.g., images and videos) from accounts we believe are connected to state-backed information operations

https://about.twitter.com/en_us/values/elections-integrity.html#data



@katestarbird
https://medium.com/@katestarbird/a-first-glimpse-through-the-data-window-onto-the-internet-research-agencys-twitter-operations-d4f0eea3f566

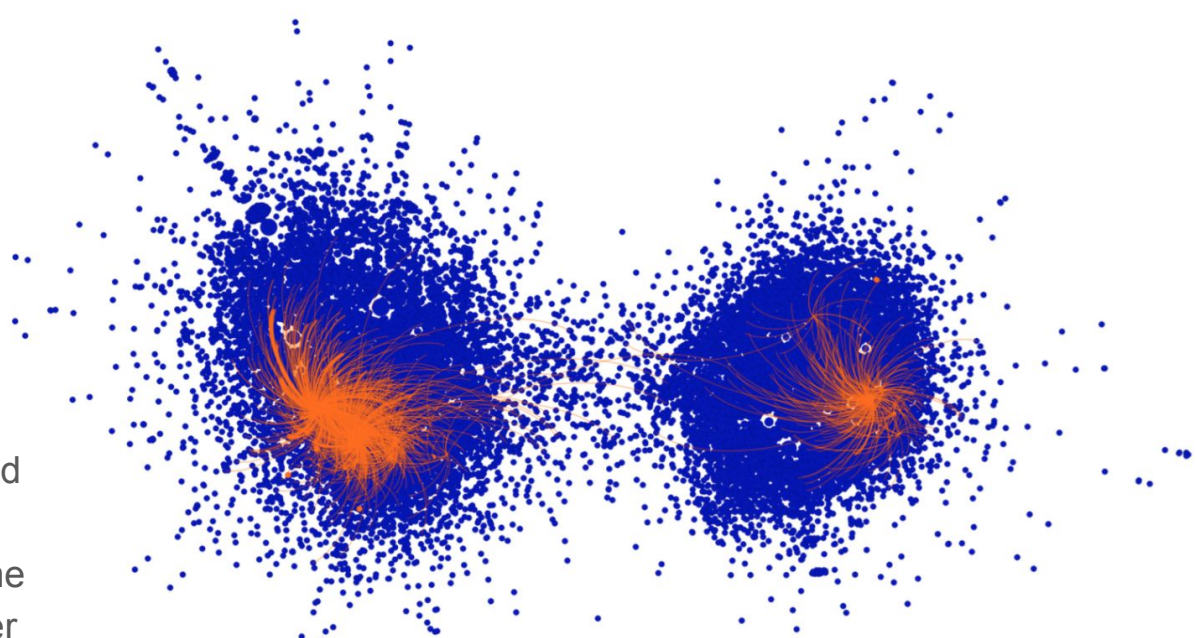

Tweets and Retweets per Week
500000
400000
300000
200000
100000
0
2010-01-01
2011-01-01
2012-01-01
2013-01-01
2014-01-01
2015-01-01
2016-01-01
2017-01-01
2018-01-01
RTs of EN tweets
English tweets



@katestarbird

https://medium.com/@katestarbird/a-first-glimpse-through-the-data-window-onto-the-internet-research-agencys-twitter-operations-d4f0eea3f566

Accounts that tend to retweet each other related to the #BlackLivesMatter Movement



https://medium.com/s/story/the-trolls-within-how-russian-information-operations-infiltrated-online-communities-691fb969b9e4

“On the political right in that conversation, IRA activity converged to support Donald Trump. On the political left in that conversation, IRA activity functioned to amplify narratives that were critical of Hillary Clinton and encouraged community members not to vote.”

https://medium.com/@katestarbird/a-first-glimpse-through-the-data-window-onto-the-internet-research-agencys-twitter-operations-d4f0eea3f566

Recap: What have learned

- Strategies like **distraction** are more common than overt strategies
- Posts cover different aspects of events
- **Timing** is important
  - Propaganda is more prevalent around specific events

Other work on these strategies

- Boyd, Ryan L., et al. "Characterizing the Internet Research Agency's Social Media Operations During the 2016 US Presidential Election using Linguistic Analyses." *PsyArXiv. October* 1 (2018).
- Spangher, Alexander, et al. "Analysis of Strategy and Spread of Russia-sponsored Content in the US in 2017." *arXiv preprint arXiv:1810.10033* (2018).
- Rozenas, Arturas, and Denis Stukal. "How Autocrats Manipulate Economic News: Evidence from Russia's State-Controlled Television." (2018).
- Paul, Christopher, and Miriam Matthews. "The Russian "firehose of falsehood" propaganda model." *Rand Corporation*(2016): 2-7.
- Munger, Kevin, et al. "Elites tweet to get feet off the streets: Measuring regime social media strategies during protest." *Political Science Research and Methods* (2018): 1-20.

...

Social Science Theory of Media Manipulation (Step 2)

Communications Theory of Media Manipulation

- Agenda setting
  - ***What*** topics are covered
- Framing
  - ***How*** topics are covered
- Priming
  - What ***effect*** reporting has on public opinion
  - “Framing works to shape and alter audience members’ interpretations and preferences through priming”

Entman’s thesis: we can use this framework to understand bias in the media

“agenda setting, framing and priming fit together as *tools of power*”

Entman, Robert M. "Framing bias: Media in the distribution of power." *Journal of communication* 57.1 (2007): 163-173.

Agenda Setting

“the media may not be successful much of the time in telling people what to think, but is stunningly successful in telling its readers what to think *about*” (Cohen, 1963)

Framing

“process of culling a few elements of perceived reality and assembling a narrative that highlights connections among them to promote a particular interpretation” [Entman, 2007]

- Word Level
  - “Estate tax” vs. “Death tax”
- Topic Level
  - Abortion is a moral issue
  - Abortion is health issue
- [This should remind you agenda setting]

Ghanem, Salma I., and Maxwell McCombs. "The convergence of agenda setting and framing." *Framing public life*. Routledge, 2001. 83-98.

Automated Analysis of Media Manipulation Strategies (Step 3)

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., & Tsvetkov, Y. (2018). Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3570-3580).

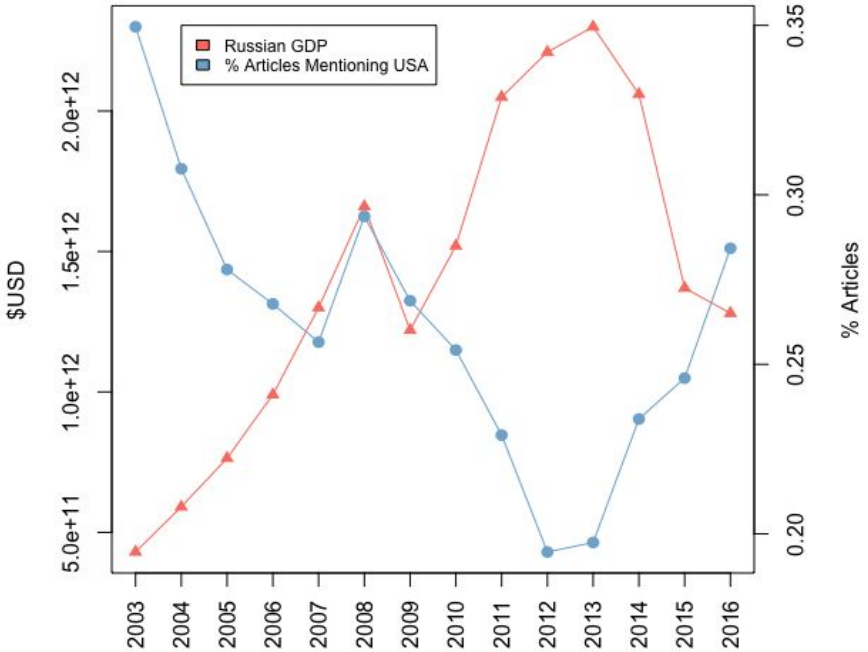
Analysis of Media Manipulation in *Izvestia*

- Data set: choose a corpus where we expect to see manipulation strategies
  - 100,000+ articles from Russian newspaper *Izvestia* (2003 - 2016)
  - Known to be heavily influenced by Russian government
- Combine **agenda-setting** with **timing** observations
  - Identify moments when we expect to see increase in manipulation strategies
  - Analyze **what** topics become more common during these times
- Analyze **framing**
  - **How** those topics are described

Benchmark against economic indicators

- Can hypothesize that we will see more more manipulation strategies during when the country is “doing poorly”
  - Government wants to distract public or deflect blame
- [Objective] measure of “doing poorly”
  - State of the economy (GDP and stock market)

Benchmark against economic indicators



State of the economy is **negatively correlated** with the about of news focused on the U.S.

	Article	**Word**
RTSI (Monthly, rubles)	-0.54	-0.52
GDP (Quarterly, USD)	-0.69	-0.65
GDP (Yearly, USD)	-0.83	-0.79

Granger Causality

$$C(w_t) = \sum_{i=1}^{m} \alpha_i (C(w_{t-i})) + \sum_{j=1}^{n} \beta_j (C(r_{t-j}))$$

- Use last month's economic state to predict this month's amount of U.S. news coverage
- Can show correlations are directed: first economy crashes, then U.S. news coverage increases

Granger Causality

$$C(w_t) = \sum_{i=1}^{m} \alpha_i (C(w_{t-i})) + \sum_{j=1}^{n} \beta_j (C(r_{t-j}))$$

	α;β	**p-value**
w_{t-1}	-0.320	0.00005
w_{t-2}	-0.301	0.0001
r_{t-1}	-0.369	0.024
r_{t-2}	-0.122	0.458

w_t frequency of U.S. mentions

r_t economic indicators

α, β coefficients learned by regression model

Framing: *How* do articles describe the U.S.?

Goals:

- Topic level
  - Identify the *primary frame* of each article
  - Identify which *frames* are present in each article
- Word Level
  - Analyze associated language

NLP approaches to framing

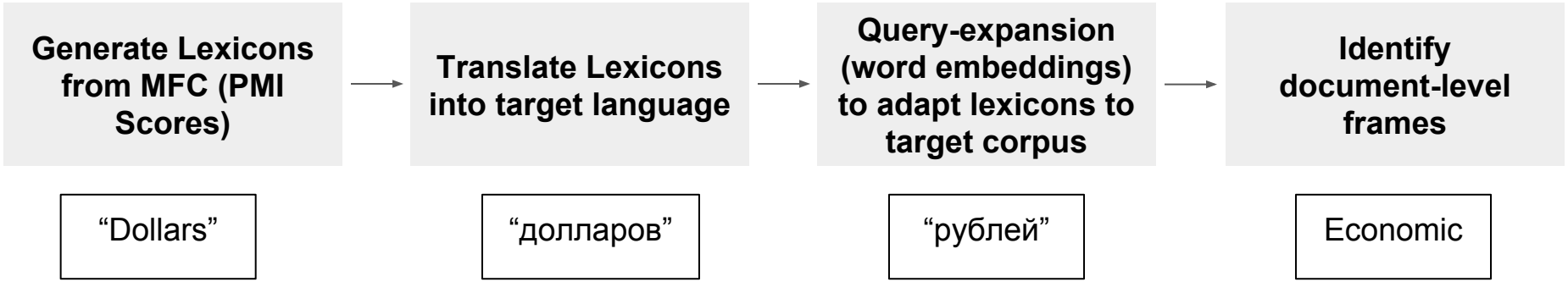
- Topic models
  - [Nguyen et al., 2013], [Boydstun et al., 2013], [Card et al., 2016]
  - *Good at identifying new frames, but difficult to interpret*
- Classifiers
  - [Baumer et al., 2015], [Ji and Smith, 2017]
  - *Easier to interpret but require annotated data*

Our approach: Media Frames Corpus

- Use **Media Frames Corpus** (MFC) for distant supervision
  - ~ 11,000 articles annotated with 14 frames
  - Spans of text are annotated with frames
  - Each article is annotated with primary frame
- Distant supervision
  - Word statistics to generate meaningful lexicons
  - Lexicons for classification

Card, Dallas, et al. "The media frames corpus: Annotations of frames across issues." *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Vol. 2. 2015.

Project MFC annotations onto *Izvestia* Corpus



English Evaluation: Frame Identification

Task: Identify the primary frame of each document (accuracy)

Ji and Smith (2017)	58.4
Card et al. (2016)	56.8
Our Model	57.3

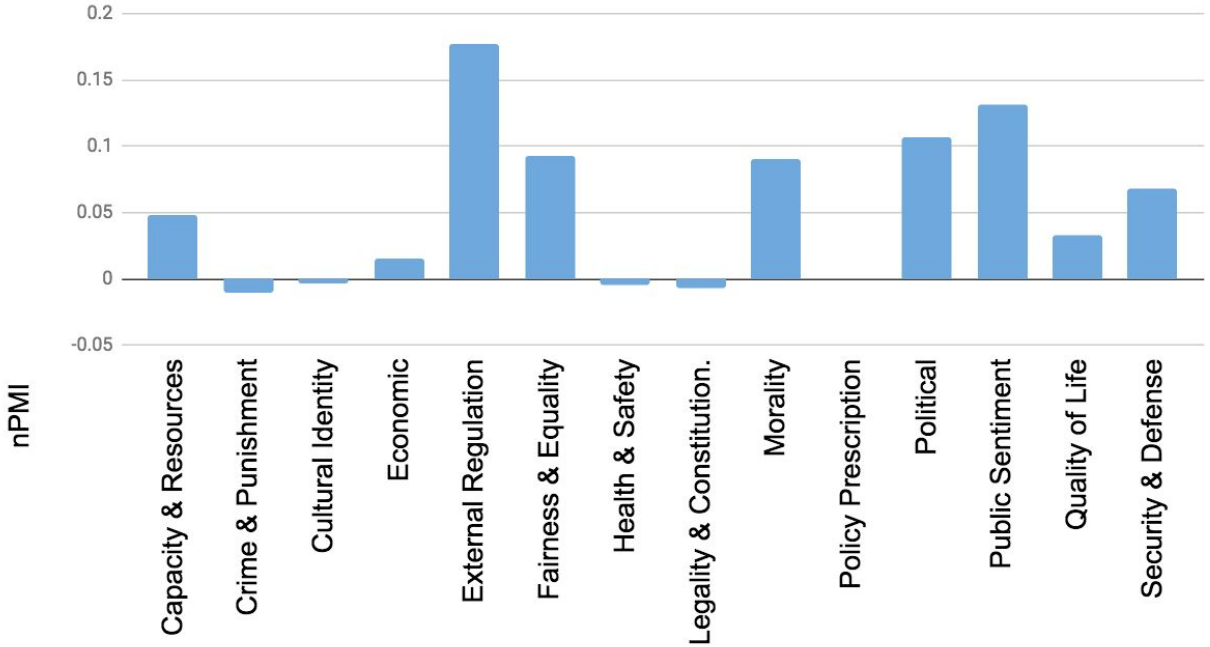
Task: Identify if frame occurs at all in each document (F1-Score)

	Ours	BOW+ Logistic Regression
Capacity & Resources	**0.53**	0.48
Crime & Punishment	**0.78**	0.76
Cultural Identity	0.57	**0.62**
Economic	**0.69**	0.67
External Regulation	0.25	**0.47**
Fairness & Equality	0.50	0.44
Health & Safety	**0.58**	0.53
Legality & Constitutionality	**0.80**	0.76
Morality	**0.31**	0.25
Policy Prescription	**0.72**	0.69
Political	**0.80**	0.77
Public Sentiment	**0.54**	0.47
Quality of Life	**0.65**	0.63
Security and Defense	0.63	0.63

Russian Evaluation: Lexicon Quality (Intruder Detection Task)

	Avg. # of annotators that correctly identified intruder
Capacity & Resources	76.67
Crime & Punishment	93.33
Cultural Identity	86.67
Economic	100
External Regulation	96.67
Fairness & Equality	40.00
Health & Safety	93.33
Legality & Constitutionality	90.00
Morality	60.00
Policy Prescription	93.33
Political	80.00
Public Sentiment	70.00
Quality of Life	80.00
Security and Defense	63.33

Which frames are most salient in U.S. focused articles?



Analysis of Specific Frames

- **Which words become more salient after downturns and less salient after upturns?**
  - Security and Defense: bombs, missiles, Guantanamo, North Korea, Iraq
- **What types of statements are said about the U.S.?**
  - “Nazi vultures...[dropped] tons of explosives into the city. The barbaric bombing of the herring caused terror and outrage all over the world”
  - “The U.S. military command no longer considers it necessary to hide its crimes, even takes pride in them”
  - "The U.S. prison in Guantanamo operates outside of all laws,"
  - “Russia will create a drones, like the U.S.”

Recap

Methods

- Use economic indicators and distantly supervised framing projects to analyze 13 years of *Izvestia* articles

Agenda-setting

- As the economy declines, U.S. news coverage increases

Framing

- New coverage of the U.S. focuses on threats to the U.S. and moral failings

Higher-level recap

- Propaganda strategies can be subtle and hard to detect
- We can identify subtle strategies by drawing from economics and political science research
  - Specifically using concepts of **agenda-setting** and **framing**
  - Economic indicators as objective benchmarks + **granger causality**
- Ethical implications
  - Identifying these strategies may encourage other actors to use them
  - The more we publicize work about how to detect strategies, the more others know how to avoid detection strategies

END