Los Alamos
NATIONAL LABORATORY

Automated Detection and Characterization of Pathological Online Behavior

Alexandra DeLucia*
Johns Hopkins University

Emma Drobina*
University of Florida

Chrysm Watson Ross
University of New Mexico

Geoffrey Fairchild, Ashlynn Daughton, **Elisabeth (Lissa) Moore**
Los Alamos National Laboratory

INDE Workshop, 8/5/2021

LA-UR-21-27759

*Authors contributed equally

Motivation

- Investigate behavioral dynamics of online communities on a large scale
- Decrease amount of human expert effort required via machine learning
  - NOTE: we never intend to completely remove humans from the loop
- Machine learning task is discovery, characterization, and understanding
- Quantitatively test sociological theories and case studies

Research Questions

- How well can communities with pathological (dangerous) behavior be identified?
- How can modern machine learning technique be used to better understand online communities?
  - Requires explainable knowledge discovery capabilities
- How can we define an online community?
  - Shared content
  - Behaviors of members (and leaders)
  - Social network characteristics
- What do disparate online communities have in common?

Data Collection

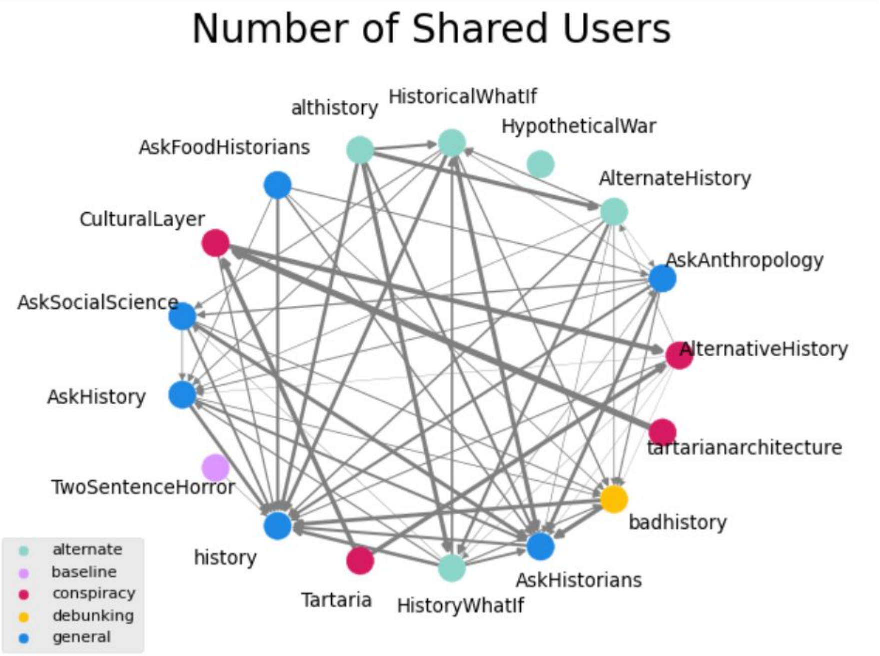
- Curated list of 15 history-related subreddits
- One baseline creative writing subreddit
- Manually categorized into community type
- All subreddit data through June 1, 2021

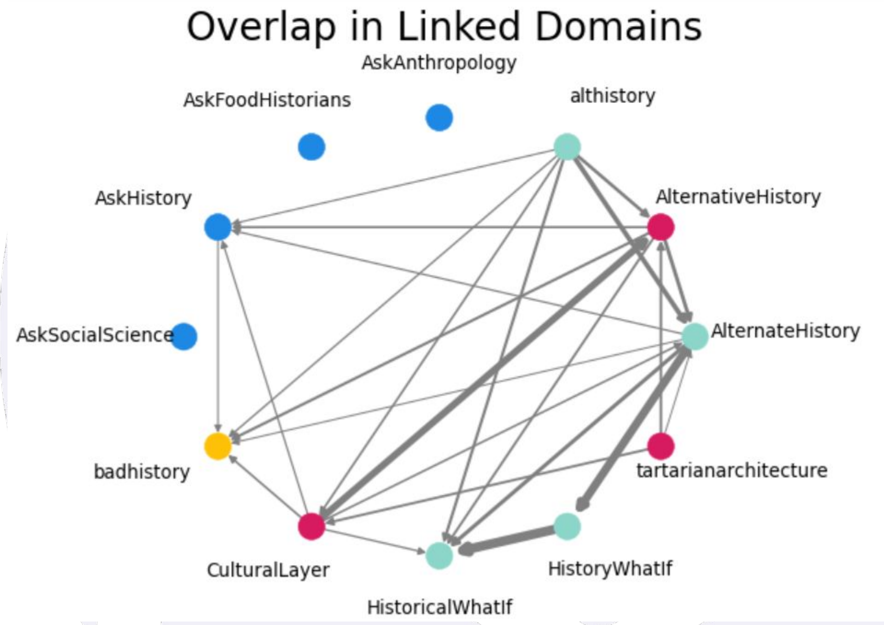
Category	Name	Founded	Subscribers	# Active	Posts
What-if	HistoricalWhatIf	2011-05-21	76,328	17,959	12,906
	althistory	2011-09-10	8,203	2,137	2,418
	HistoryWhatIf	2014-12-26	81,106	16,941	28,068
	AlternateHistory	2010-06-20	67,693	14,185	13,028
	HypotheticalWar	2013-07-06	396	58	53
Conspiracy	CulturalLayer	2017-09-10	38,884	3,964	2,454
	tartarianarchitecture	2018-12-18	3,826	575	1,727
	AlternativeHistory	2008-08-03	123,633	10,577	6,983
	Tartaria	2018-12-26	9,733	1,551	1,211
Debunking	badhistory	2013-03-19	248,373	26,339	6,345
General	AskHistory	2011-01-20	78,239	24,052	19,198
	history	2008-01-25	15,887,782	395,094	145,369
	AskSocialScience	2011-07-09	101,227	21,601	17,599
	AskAnthropology	2013-03-10	121,382	17,795	11,107
	AskFoodHistorians	2013-01-12	40,202	4,251	1,008
Baseline	TwoSentenceHorror	2014-03-05	656,864	25,620	66,588

Collected with https://github.com/AADeLucia/retriever

Exploratory Data Analysis: Content and User Overlap

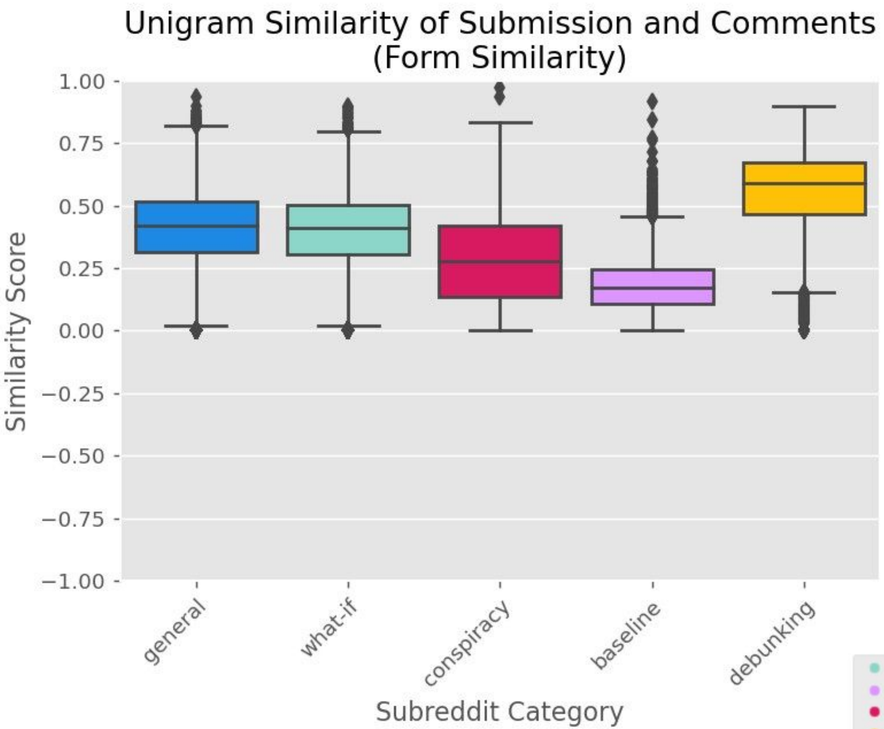
- Conspiracy-based communities may be more insular echo chambers

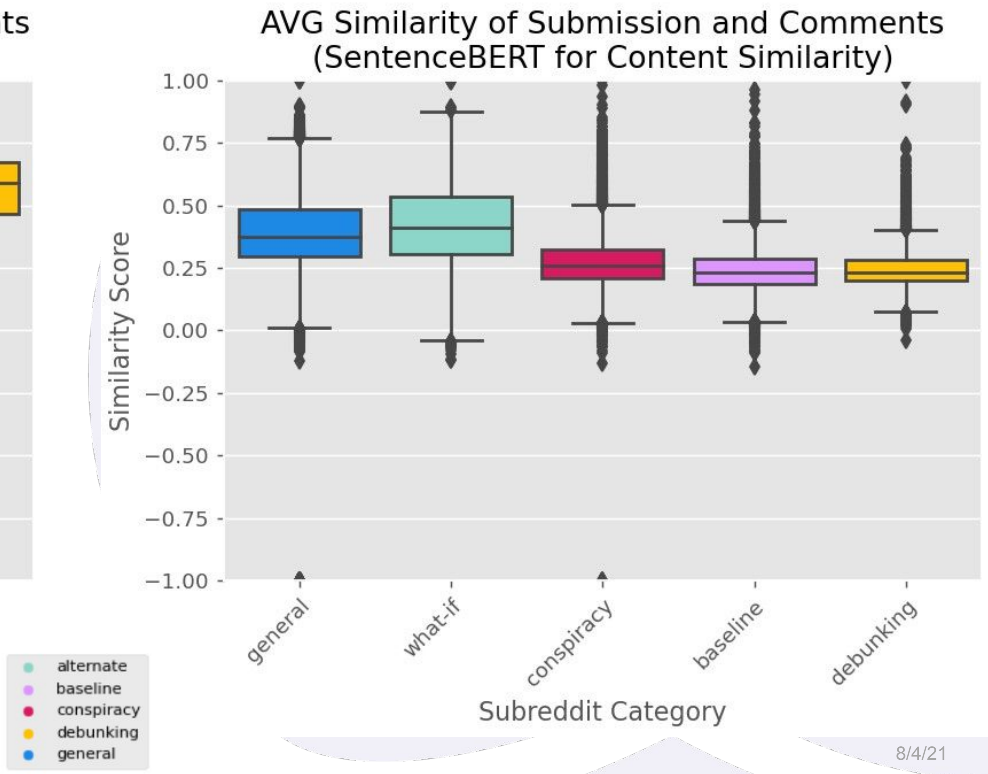




Exploratory Data Analysis: Intra-Community Similarity

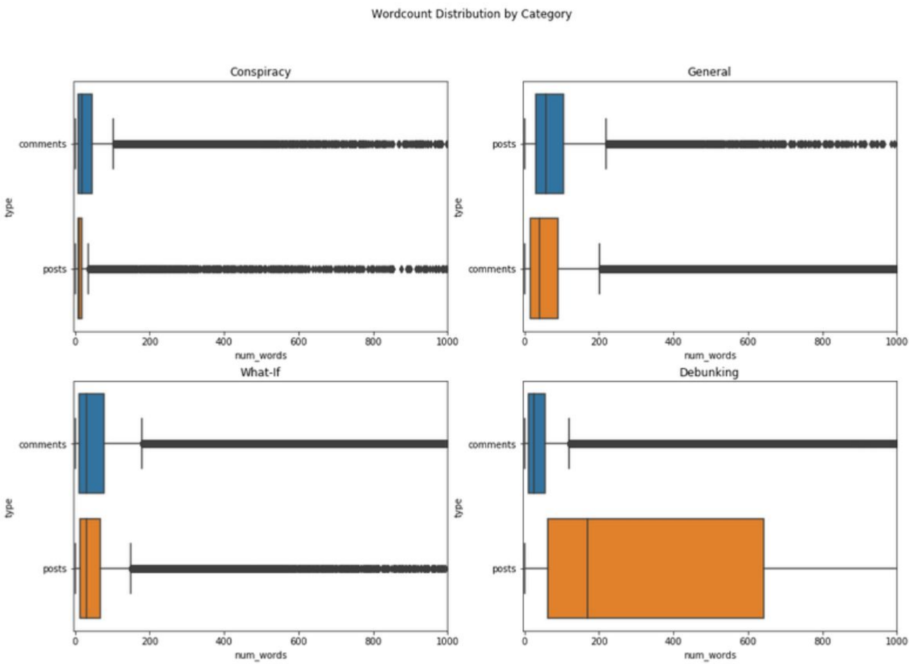
- Community types show statistically significant differences in characteristics

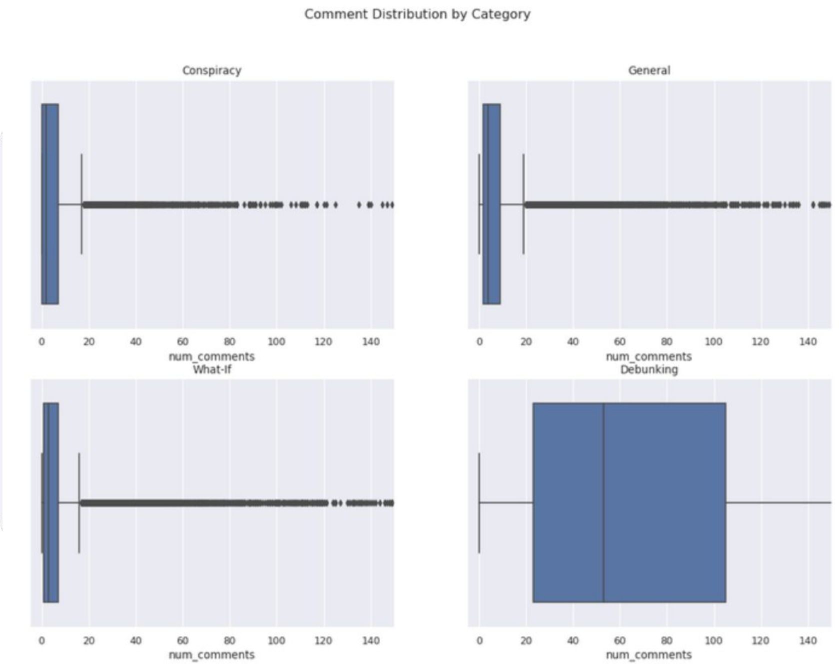




Exploratory Data Analysis: Post & Comment Metadata

- Distinct signals also appear in metadata such as word and per-post comments





Task: URL Domain Type Prediction

Goal: Understand how users introduce and discuss types of websites

1. Collect URLs from metadata and raw text and parse domain

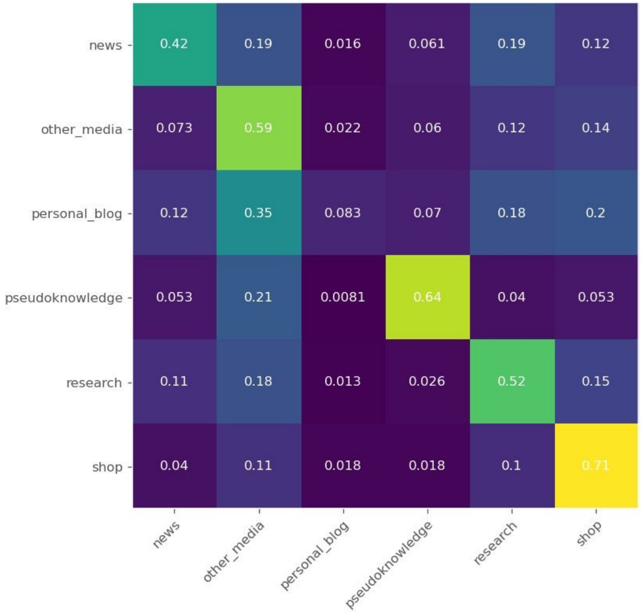
2. Manually label domains into one of 13 categories
    - Modified from (Introne, 2018)
    - Pseudoknowledge, reference, science, shop, other media, etc

3. Supervised learning of domain category given a variety of post/comment features

Task: URL Domain Type Prediction

- Train/Validation: 38,133 Test: 9,534
- Group historical, reference, science, academic together for better performance

AVG Train F1 (n=100)	AVG Test F1
0.61 (0.027)	0.6 (0.027)



Domain Category	Test Class	F1
research	59.47%	0.71 (0.04)
other_media	23.94%	0.53 (0.015)
news	6.50%	0.29 (0.017)
personal_blog	4.80%	0.16 (0.025)
shop	3.42%	0.33 (0.035)
pseudoknowledge	1.88%	0.35 (0.054)

Los Alamos NATIONAL LABORATORY

Task: URL Domain Type Prediction

- Binary: Research vs. Pseudoknowledge
- Identifying one specific type of website domain is slightly easier task
- Low Pseudoknowledge precision is due to confusion with Wikipedia

Test (n=100)	F1	Precision	Recall
Pseudoknowledge (3.07 %)	0.38 (0.079)	0.26 (0.073)	0.81 (0.053)
Research (96.93%)	0.95 (0.026)	0.99 (0.0016)	0.91 (0.047)

Task: URL Domain Type Prediction

- Multiclass prediction of domain type is difficult for automated methods
- “Research” vs “Pseudoknowledge” is also difficult
- Exploratory clustering could be better suited to this task
- Issues
  - Different communities may use the same type of domain in different
  - Labels at the domain-level are coarse and do not capture the content on the specific webpage

Task: Explainable Clustering

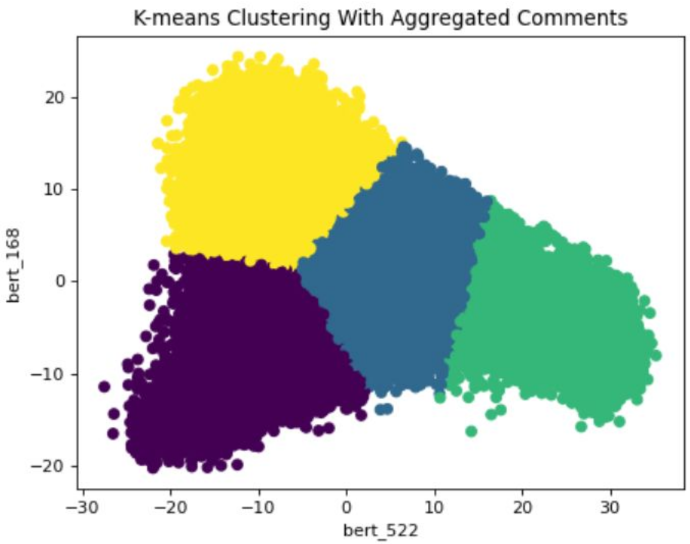
1. Comments and their metadata are aggregated (mean, median, min, max, standard deviation) and combined with posts and post metadata

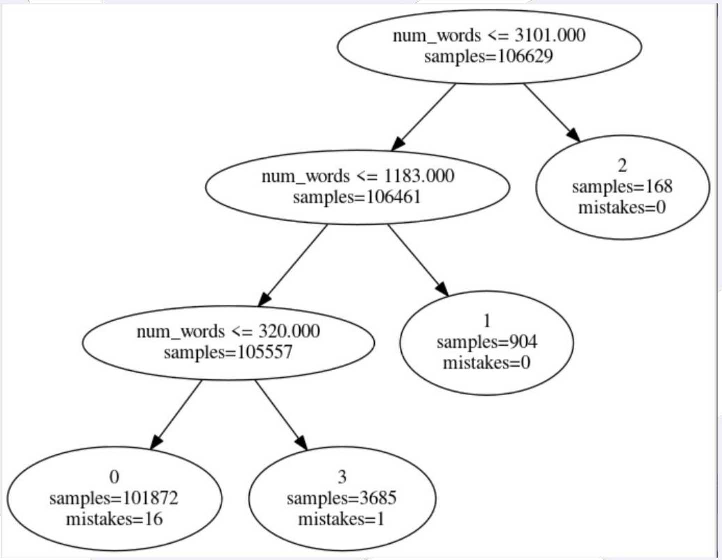
2. Current state-of-the-art for explainable clustering: combine k-means and decision trees

3. Incorporate supervised learning (using our category assignments as the ground truth) to get a better sense of the features

Task: Explainable Clustering

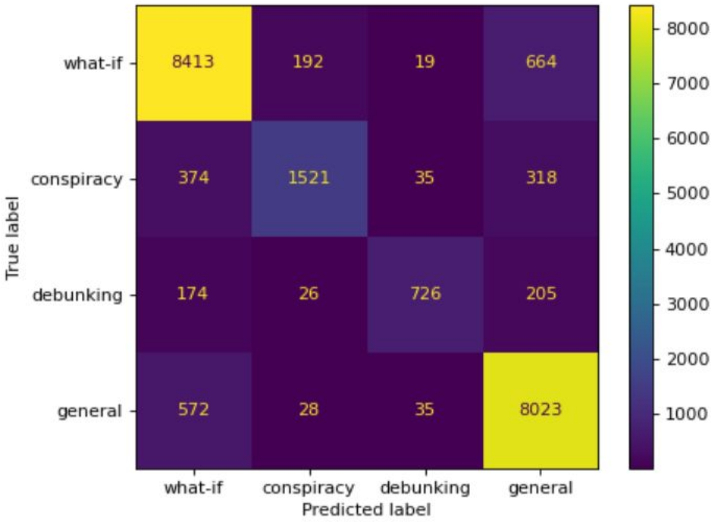
- Clusters do not separate well
  - Rand score (cluster accuracy) of 61.0%
  - Explanations only focus on metadata

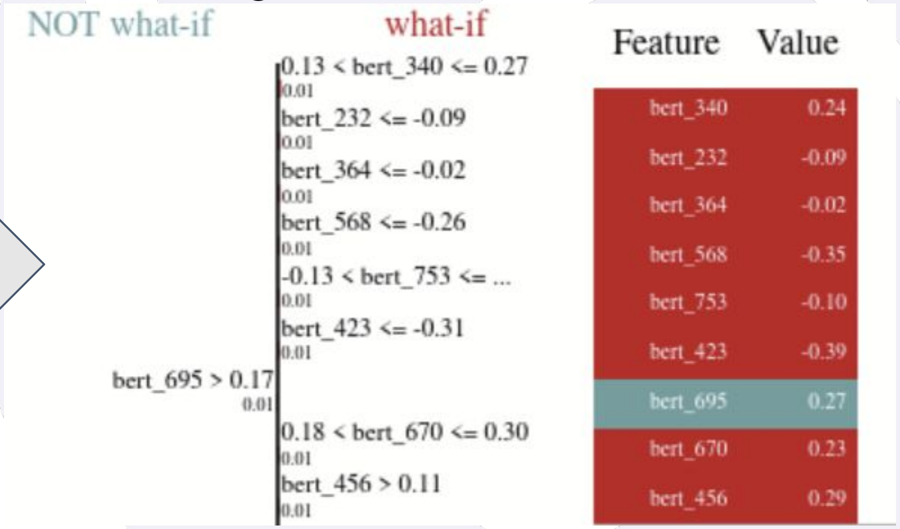




Task: Explainable Clustering

- Supervised methods find significantly higher accuracy
  - 87.6% for random forests
  - 84.0% for neural network (by-class accuracy is more even)
- Classification depends primarily on text embeddings





Task: Shared Link Validity Detection

- Join with auxiliary dataset from Newsguard
- Represent subreddits of interest as dynamic hypergraphs
- Preliminary results show vast majority of shared links are from reputable sites
  - Cross-posts and re-posts are rare and hard to detect
- Implication is that majority of misinformation arises purely from intra-community chatter and speculation

Summary (Challenge Problems)

- Need for new, flexible methods tailored to knowledge discovery tasks in online platform data
  - Unsupervised methods
  - Explainability techniques

- Need for nuanced methods for detecting similar content, and dynamics surrounding content, that doesn't rely on shared links

Thank you!

lissa@lanl.gov