

Automated Detection and Characterization of Pathological Online Behavior

Alexandra DeLucia*
Johns Hopkins University

Emma Drobina*
University of Florida

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

AML Student Symposium, 8/12/2021
LA-UR-21-28109

Understanding Online Communities

- Previous work in studying misinformation and conspiracy groups performed mostly on a small scale
- Little work on automating these types of case studies
- Understand the communities through modeling
 - Shared content, behaviors of members (and leaders), social network
- Focus on Reddit, one of the largest online communities
 - Collection of history subreddits across the spectrum
 - Manually categorized as “general”, “conspiracy”, “what-if”, and “debunking”



Understanding Online Communities

- Previous work in studying misinformation and conspiracy groups performed mostly on a small scale
- Little work on automating these types of case studies
- Understand the communities through modeling
 - **Shared content**, behaviors of members (and leaders), social network
- Focus on Reddit, one of the largest online communities
 - Collection of history subreddits across the spectrum
 - Manually categorized as “general”, “conspiracy”, “what-if”, and “debunking”



Community Shared Content

 **r/history** · Posted by  2 days ago   2

Why Richard the Lionheart so romanticize

Discussion/Question

I personally have always been fascinated by Richard I. I'm inspired by his leadership and military success, especially to the British. Considering he rebelled against his father, he spent months in England, used it to build his army, pay for his wars, and speak English. Even his moniker was originally French (Coeur de Lion).

Is there a culture in UK where courage and military competence are very admirable traits; but as great as a military man Richard I was, he was also a man of peace.

 348 Comments  Award  Share  Save 

Text post from r/history

Raw text

 **r/Tartaria** · Posted by  2 days ago

Wow, the research in this video is over the top. A must watch.

youtu.be/mBNsx8...







 21 Comments  Award  Share  Save  Hide  Report 94% Upvoted

YouTube post from r/Tartaria

Submission with link




If Japan had refused to surrender after the second nuking AND launched Operation Cherry Blossoms... yeah that's a nightmare situation. One thing for certain is that the entirety of the Korean peninsula would have been occupied by the Soviets while the pacific war rages on.


 6   Reply Give Award Share Report Save

 -1d

Do you have a source about the radio communication?




That sounds really interesting so I'd like to read more.

 1   Reply Give Award Share Report Save

 OP - 1d

https://en.wikipedia.org/wiki/Atomic_bombings_of_Hiroshima_and_Nagasaki#Events_of_7-9_August

On 7 August, a day after Hiroshima was destroyed, Dr. Yoshio Nishina and other atomic physicists arrived at the city, and carefully examined the damage. They then went back to Tokyo and told the cabinet that Hiroshima was indeed destroyed by a nuclear weapon. Admiral Soemu Toyoda, the Chief of the Naval General Staff, estimated that no more than one or two additional bombs could be readied, so they decided to endure the remaining attacks, acknowledging "there would be more destruction but the war would go on".[184] American Magic codebreakers intercepted the cabinet's messages.[185]

 3   Reply Give Award Share Report Save

Wiki link in comments of r/HistoryWhatIf

Comment with link

Community Shared Content

 **r/history** · Posted by  2 days ago   2

Why Richard the Lionheart so romanticize

Discussion/Question

I personally have always been fascinated by Richard I. I'm inspired by his military achievements, but I haven't really come across anything that was especially to the British. Considering he rebelled against his father, spent months in England, used it to build his army, pay for his wars, and speak English. Even his moniker was originally French (Coeur de Lion).

Is there a culture in UK where courage and military competence are very admirable traits; but as great as a military man Richard I was, he was also a man of letters.

348 Comments  Award  Share  Save 

Text post from r/history

Raw text

 **r/Tartaria** · Posted by  2 days ago

Wow, the research in this video is over the top. A must watch.

youtu.be/mBNsx8...






21 Comments  Award  Share  Save  Hide  Report 94% Upvoted

YouTube post from r/Tartaria

Submission with link



If Japan had refused to surrender after the second nuking AND launched Operation Cherry Blossoms... yeah that's a nightmare situation. One thing for certain is that the entirety of the Korean peninsula would have been occupied by the Soviets while the pacific war rages on.


6   Reply Give Award Share Report Save

 -1d

Do you have a source about the radio communication?



That sounds really interesting so I'd like to read more.

1   Reply Give Award Share Report Save

 OP - 1d

https://en.wikipedia.org/wiki/Atomic_bombings_of_Hiroshima_and_Nagasaki#Events_of_7-9_August

On 7 August, a day after Hiroshima was destroyed, Dr. Yoshio Nishina and other atomic physicists arrived at the city, and carefully examined the damage. They then went back to Tokyo and told the cabinet that Hiroshima was indeed destroyed by a nuclear weapon. Admiral Soemu Toyoda, the Chief of the Naval General Staff, estimated that no more than one or two additional bombs could be readied, so they decided to endure the remaining attacks, acknowledging "there would be more destruction but the war would go on".[184] American Magic codebreakers intercepted the cabinet's messages.[185]

3   Reply Give Award Share Report Save

Wiki link in comments of r/HistoryWhatIf

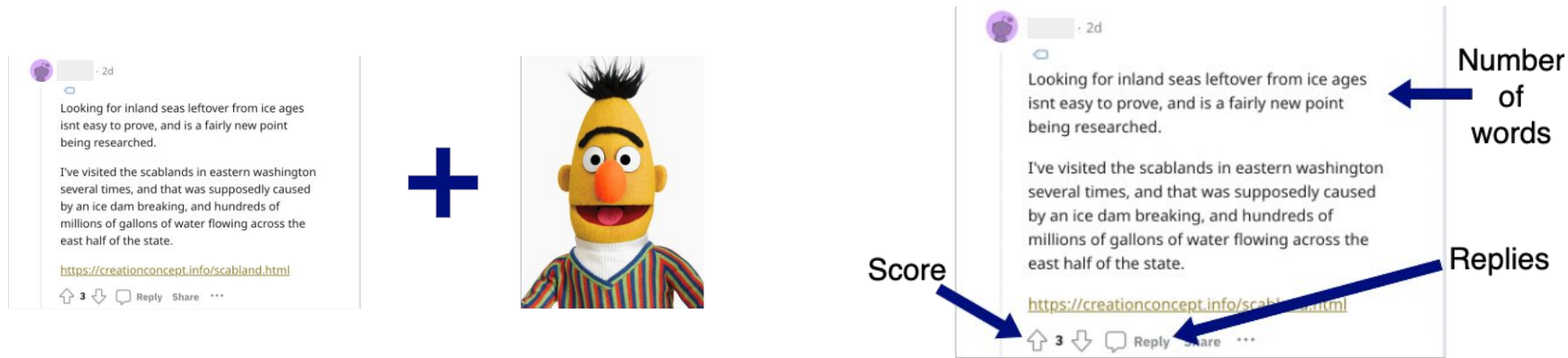
Comment with link

How do users introduce and discuss types of websites?

Approach

1. Gather URLs from **posts and comments**
2. **Manually** categorize URLs by domain
 - Mix of regex and manual investigation
3. Represent each URL sample with features from the post/comment
 - Features: **interactions** with the post/comment and **context** (BERT)
4. Train a model on the labelled data to **predict domain category**
 - Fully connected neural net
5. Explain model's decisions
 - Goal: gain **insights** into how types of domains are received/described by users

Domain Prediction Features



BERT representation of post

Content

Inclusion of raw metadata fields

Interactions / behavior

Domain Prediction Model

- Fully connected neural net with 3 hidden layers and ReLU activation
 - Layer sizes = {512, 256, 128, 32}
- Adam optimizer with 0.001 learning rate
 - L2 regularization = 0.001
 - BCE loss with balanced class weights
- 100 epochs with batch size 200
- Early stopping with patience of 10 epochs
 - Separate 10% of training set for validation
 - Stopped if validation score does not improve by 0.05 F1



Domain Categories

- Domain categories modified from (Introne et al, 2018) for Reddit

Category	Description	Example Domains
Personal blog	Website dedicated to the opinions and interest of a few people.	blogspot.com, mikedashhistory.com, ehmanblog.org
News	News websites. Not labelled for bias.	apnews.com, bloomberg.com, foxnews.com, timesofmalta.com
Academic*	University-hosted websites.	wits.ac.za, cam.ac.uk, cardiff.ac.uk
Reference*	General reference sites.	archives.gov, bibliotecapleyades.net, mises.org, visualcapitalist.com
Wikipedia*	All Wikipedia-based domains.	wikipedia.org, wikisource.org, wikiwand.com
Reddit-reference	Internal references to other subreddits (crossposts).	self.*, reddit.com
Historical*	Sites dedicated to spreading historical knowledge.	ancient-archeology.com, iea.org.uk, nps.gov, smithsonianmag.com
Science*	Published and pre-print paper websites and credible pop science magazines.	academia.edu, arxiv.org, researchgate.net
Pseudoknowledge	Blogs and websites that promote conspiracy theories and pseudoknowledge/science.	abovetopsecret.com, alienpolicy.com, stolenhistory.org, mysteriousuniverse.org
Shop	Online stores.	kickstarter.com, subterraneanpress.com, etsy.com
Other media	General multi-media that cannot be categorized elsewhere. Art, image-hosting sites, fandom, tourism, social media, forums, games, and general human interest sites.	wikia.com, facebook.com, theonion.com
YouTube	Videos hosted on YouTube.	youtube.com,youtu.be
Reddit-media	Videos and images uploaded directly to Reddit.	reddituploads.com, redd.it
Unknown	Sites that cannot be categorized because they are no longer hosted.	antimachiavel.com, cloudriot.com, 123flatshare.com

Research


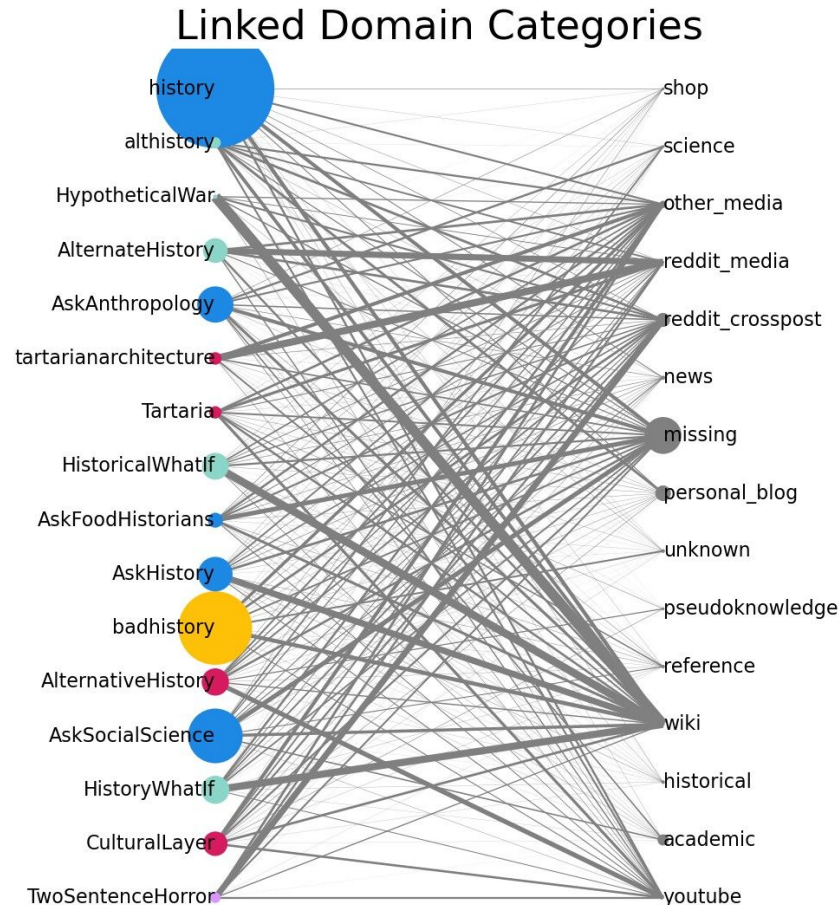


TABLE II: The URL domain categories with their descriptions and examples. The categories are a modified version of those in [5]. YouTube and Reddit-media are separated from Other media because they were the largest subsets. Categories marked with an asterisk (*) are further grouped into “Research.”

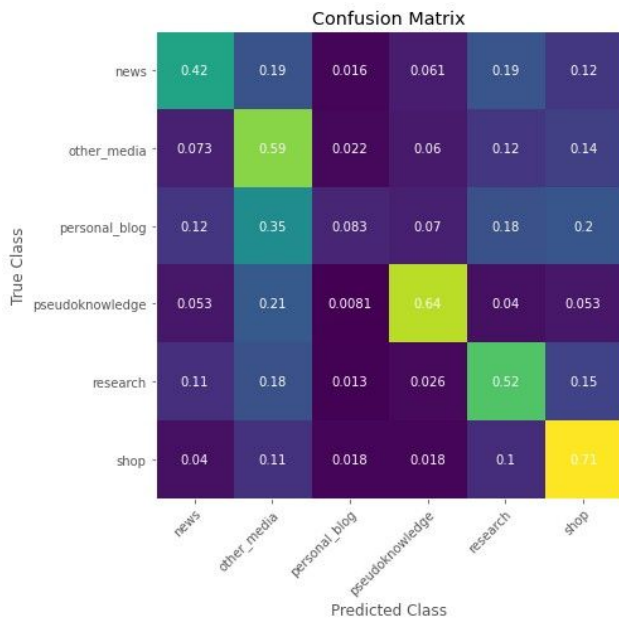
Domain Category Distribution

- Most posts and comments with links are from r/history and r/badhistory
- Pseudoknowledge links are not common
- Wikipedia and YouTube are commonly used across all subreddits



Results: Supervised

- Variable per-class performance



Domain Category	Class Dist	Test F1
All	100%	0.6 (0.027)
All (Random)	100%	0.44 (0.0032)
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)

Results: Supervised Research vs Pseudoknowledge

- Low Pseudoknowledge F1 is due to very low precision

Test (n=100)	F1	Precision	Recall
Pseudoknowledge (2.83%)	0.37 (0.076)	0.24 (0.069)	0.82 (0.052)
Research (97.17%)	0.95 (0.02)	0.99 (0.0015)	0.92 (0.036)

What indicators of pseudoknowledge is the model identifying?

Results: Supervised Research vs Pseudoknowledge

- Grouping of sensational content under pseudoknowledge

“Giants, demons having sex
with humans, it's all here:
The Book of Enoch”

“The Great Pyramid
Experiment: Measuring the
Sonic Capabilities of the
Dead-end Passage”

“Spy Satellites Reveal
Ancient Lost Empires in
Afghanistan”

“There’s No Scientific Basis
for Race—It's a Made-Up
Label”

Results: Supervised Research vs Pseudoknowledge

- Grouping of sensational content under pseudoknowledge



“Giants, demons having sex with humans, it's all here: The Book of Enoch”

“The Great Pyramid Experiment: Measuring the Sonic Capabilities of the Dead-end Passage”



LIVESCIENCE

“Spy Satellites Reveal Ancient Lost Empires in Afghanistan”

NATIONAL GEOGRAPHIC

“There’s No Scientific Basis for Race—It’s a Made-Up Label”

Results: Supervised Research vs Pseudoknowledge

- Similar use of PK and research websites

“If you read through this you will find a reference to the planet experiencing something of catastrophic proportions. The content of your post and this one document may have clues to what actually happened.”



“The official reason cited tends to be that the embargo is in place to punish Cuba for human rights violations. However, this seems to be pretty widely accepted as merely a pretext In 2008, a PAC on defending the embargo spent over a million dollars on elections. Source here .”

AlterNet

Why is this task difficult?

- Users can treat and discuss different types of websites the same
 - As a reference, as a dispute, for entertainment, etc.
 - Model can identify patterns in *how* users discuss citations
- Labels at the domain-level are distant and do not capture the content on the specific webpage
 - i.e. news on a primarily entertainment site is still “other media”
- Model has no access to web page content
 - Sometimes has access to article title
- Websites can discuss the same topics across categories
 - “Aliens exist” vs “looking for life on Mars”

Summary

- Automate the categorization of shared URLs on social media using only the context of the link
 - Created a dataset of 19K manually annotated domains (with ~850 regex)
- Difficult task for prediction models, even in the binary setting
 - Multiclass: 0.6 F1
 - Pseudoknowledge vs Research: 0.37 F1 vs 0.95 F1
- Model discovered patterns in the way links are discussed and used

Future Directions

1. Need for research focusing on automating manual social science analyses
 - Address challenges in scaling from small case studies to larger communities
2. Need for nuanced methods for detecting similar content, and dynamics surrounding content, that doesn't rely on shared links
 - Preliminary results show vast majority of shared links are from reputable sites (not pseudoknowledge)
 - Implication is that majority of misinformation arises purely from intra-community chatter and speculation

Thank you for your time. Questions?

Automate Analysis of Shared Information

- Manual analysis of outside information in social sciences

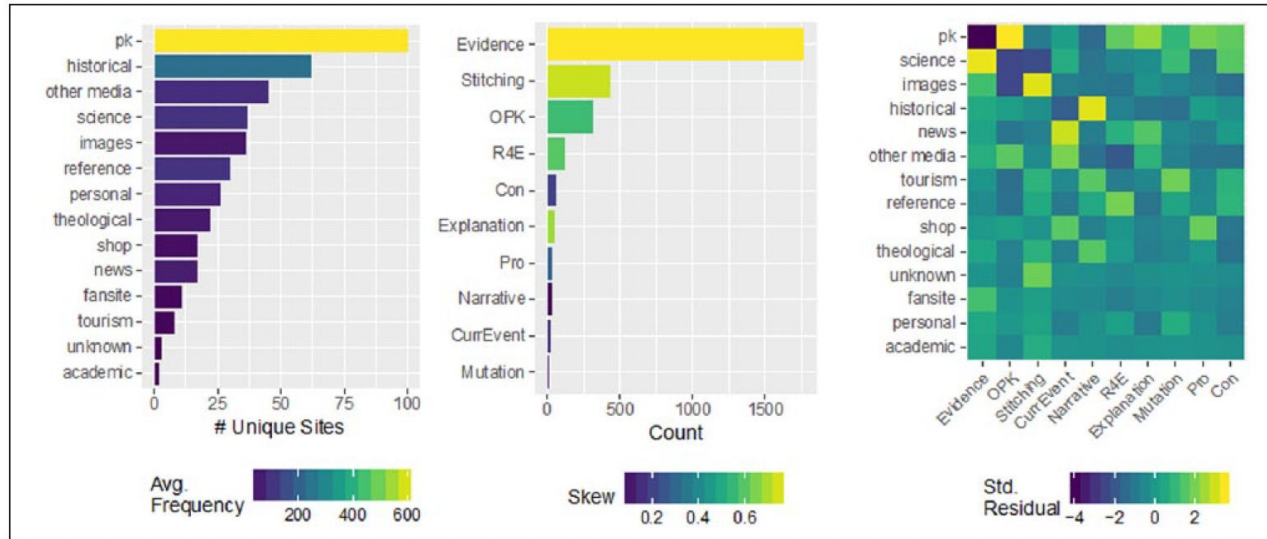


Figure 3. Distribution of the codes: (a) distribution of resource types over 415 unique domains, (b) distribution of discourse codes across 2,484 links, and (c) correlations between resource types and discourse codes, obtained from residuals of a chi-square analysis. Raw counts in each cell are provided in Appendix C.

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

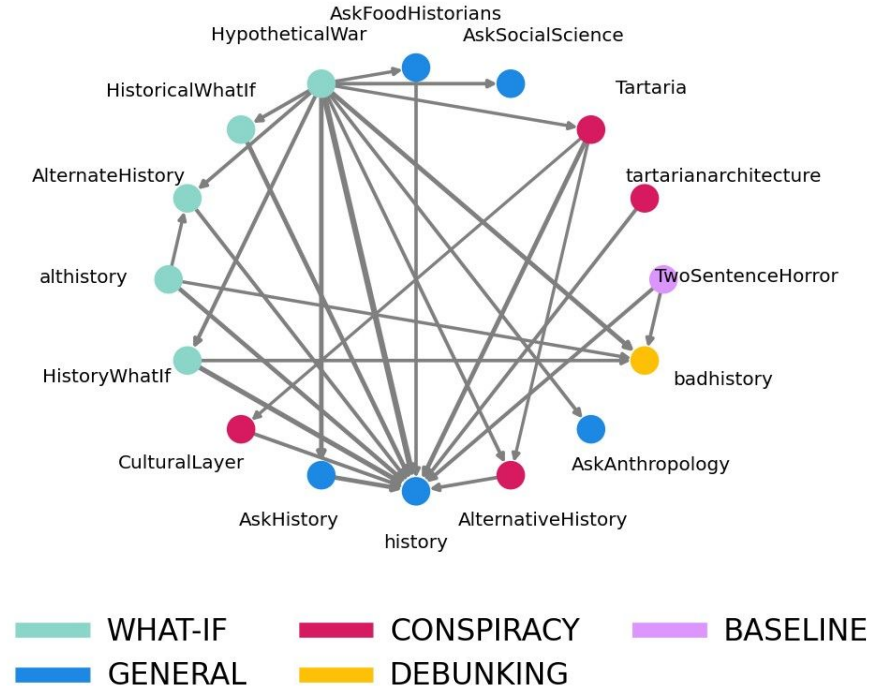


Category	Name	Founded	Subscribers	# Active	Posts
What-if	HistoricalWhatIf	2011-05-21	76,328	17,959	12,906
	althistory	2011-09-10	8,203	2,137	2,418
	HistoryWhatIf	2014-12-26	81,106	16,941	28,068
	AlternateHistory	2010-06-20	67,693	14,185	13,028
	HypotheticalWar	2013-07-06	396	58	53
Conspiracy	CulturalLayer	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

Domain Overlap between Subreddits

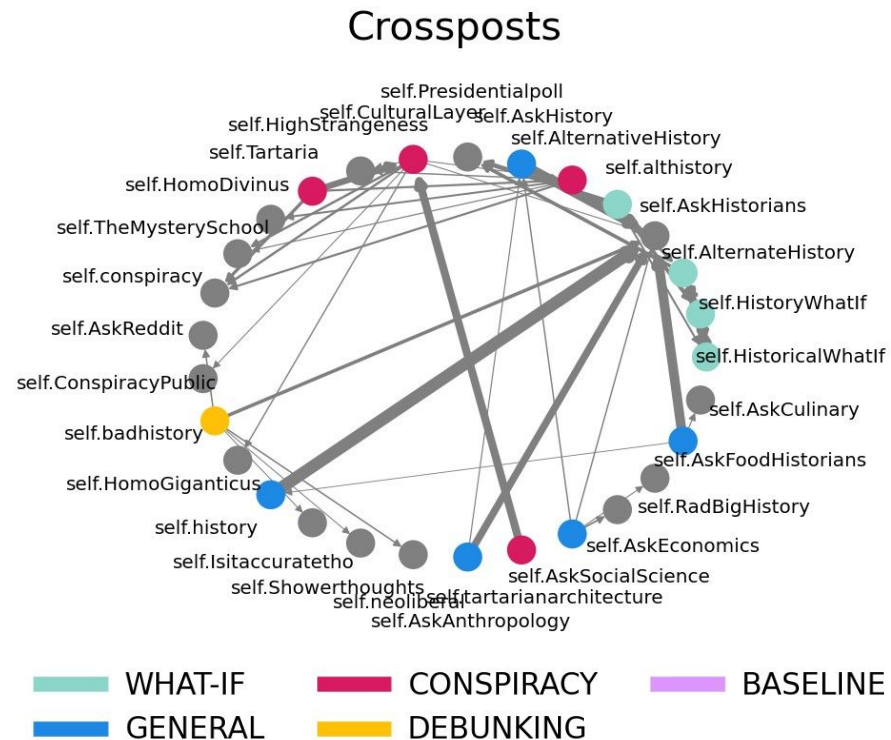
- Overlap in domains can indicate overlap in discussed content
- Some domains are used universally and more insight can be gained from deeper analysis
 - e.g. Wikipedia

Overlap in Linked Domains



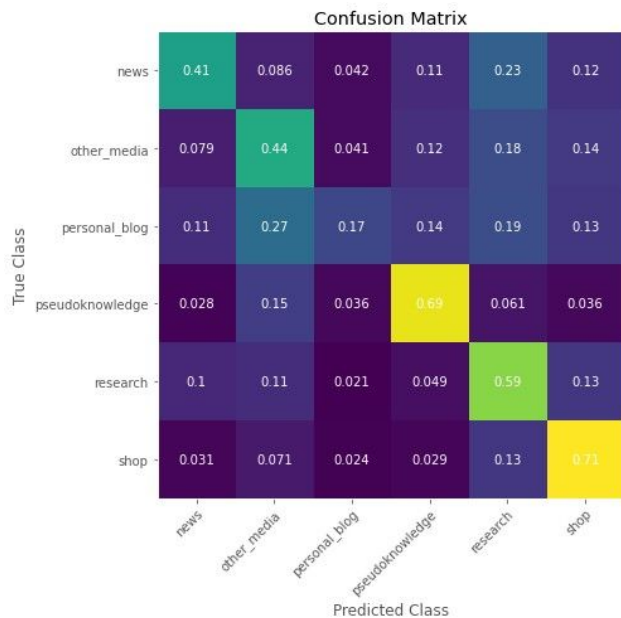
Popular Crossposts per Subreddit

- Conspiracy groups draw from each other
- Rate of crossposts can indicate greater community overlaps



Results: Supervised (not OvR)

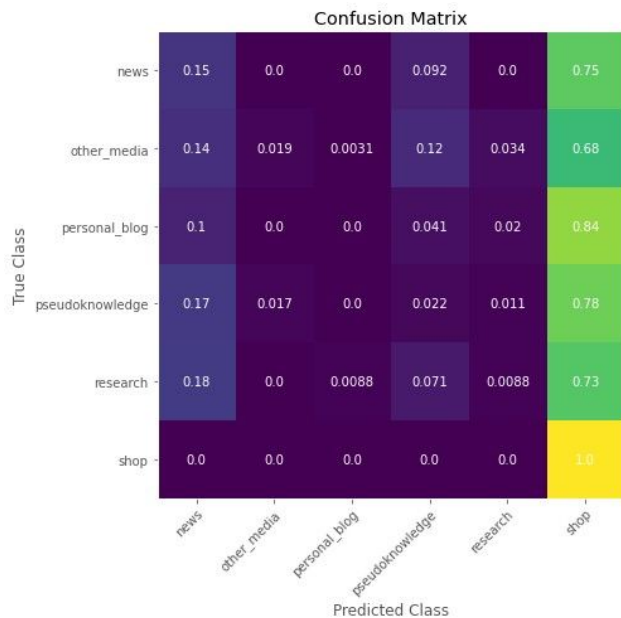
- Regular multiclass performs worse than One vs Rest



Domain Category	Class Dist	Test F1
All	100%	0.53 (0.05)
research	61.38%	0.63 (0.076)
other_media	22.71%	0.47 (0.034)
news	6.19%	0.27 (0.022)
personal_blog	4.68%	0.18 (0.019)
shop	3.25%	0.24 (0.051)
pseudoknowledge	1.79%	0.31 (0.049)

Results: Supervised w/ Aggregated Posts+Comments

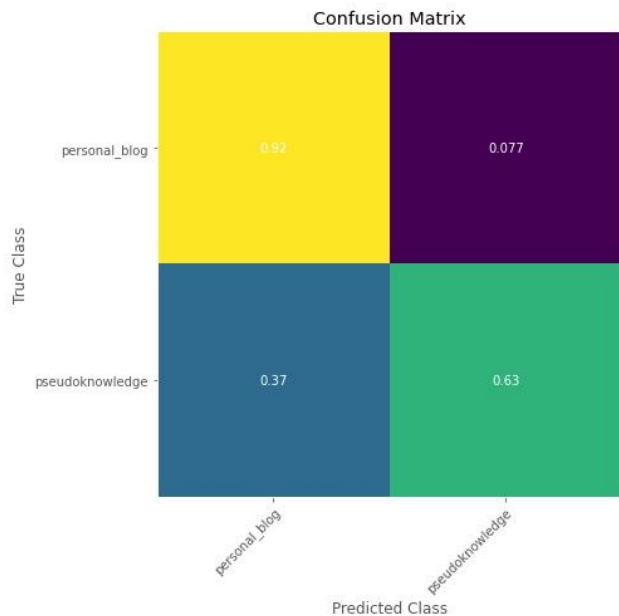
- Large performance drop due to smaller dataset



Domain Category	Class Dist	Test F1
All	100%	0.1 (0.042)
research	59.47%	0.12 (0.085)
other_media	23.94%	0.14 (0.1)
news	6.50%	0.06 (0.055)
personal_blog	4.80%	0.044
shop	3.42%	0.013 (0.012)
pseudoknowledge	1.88%	0.08 (0.12)

Results: Supervised Personal blog vs Pseudoknowledge

- Performs much better than PK vs Research despite class imbalance



Domain Category	Class Dist	Test F1
All	100%	0.81 (0.031)
personal_blog	72.37%	0.86 (0.034)
pseudoknowledge	27.63%	0.68 (0.029)