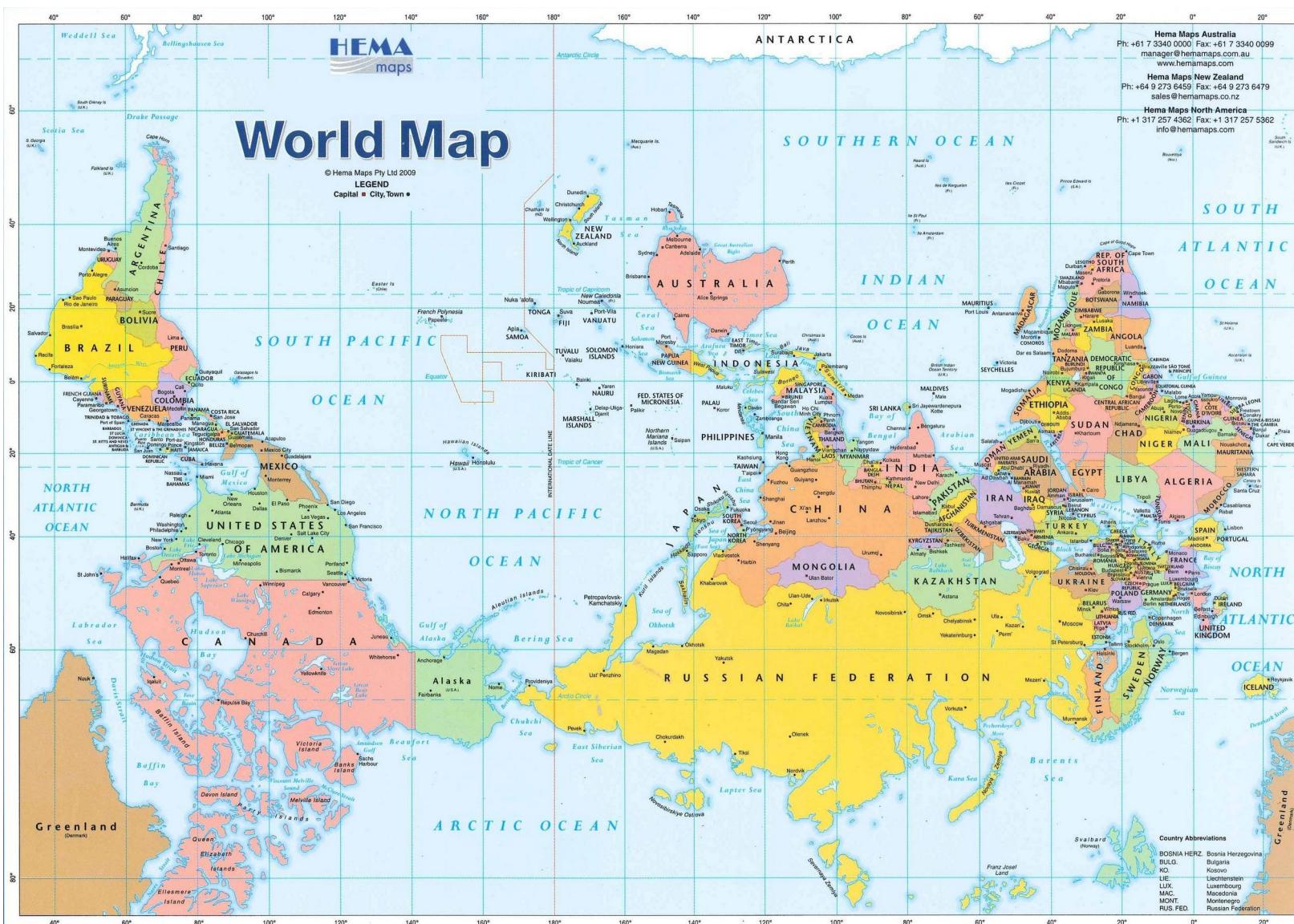


Mapping Online Problematic Content: Mixing Qualitative Approaches with State-of-the-art Machine Learning



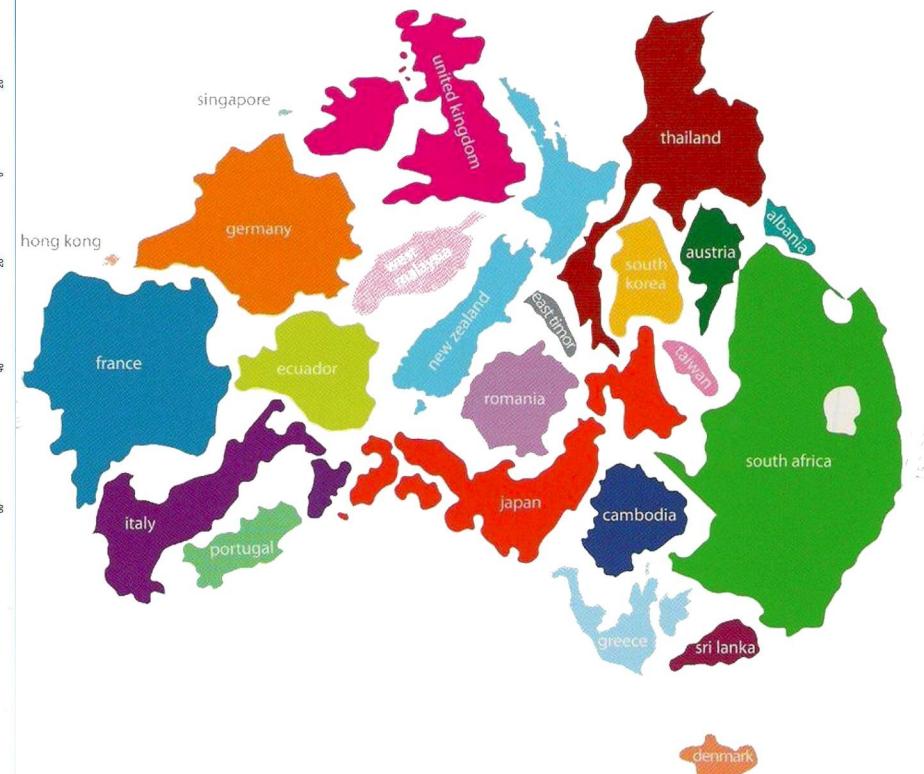
Marian-Andrei Rizoiu | Behavioural Data Science
Marian-Andrei.Rizoiu@uts.edu.au
<https://www.behavioral-ds.science>

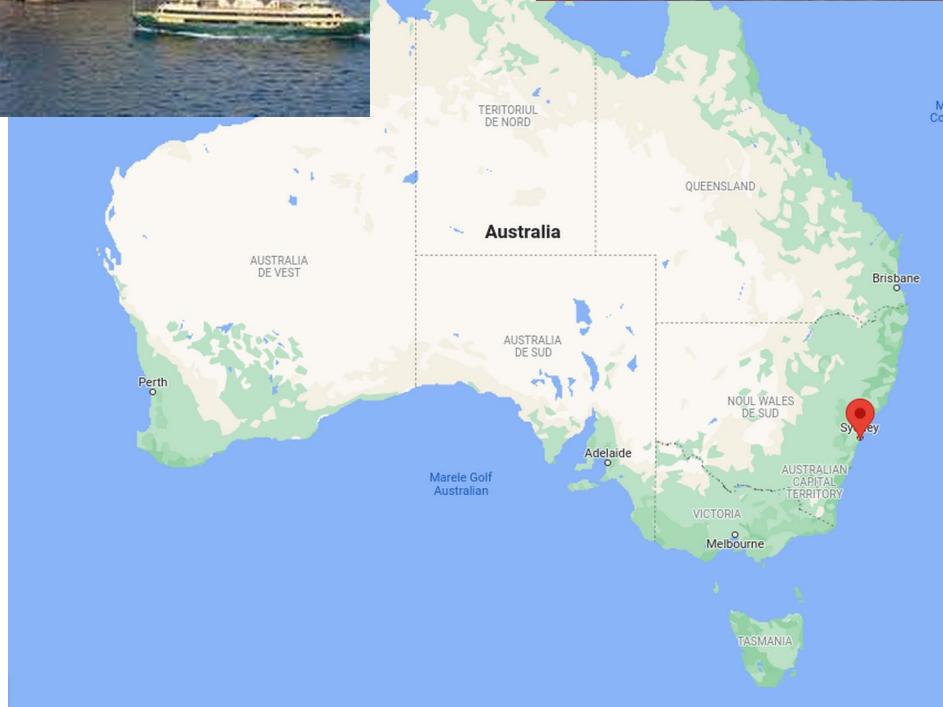
Australia



Australia is a big BIG place

The world map, according to Australians





Located in Sydney, Australia



A city campus, iconic brutalist style
blended with modern buildings

The research group



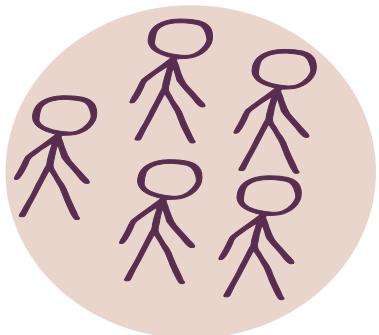
Behavioral
Data Science

1 PostDoc, 6 PhD, 3 Masters, 1 assistant prof.



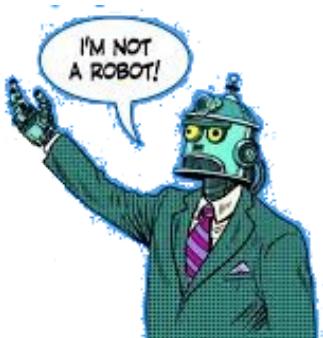
The Behavioral Data Science

1.



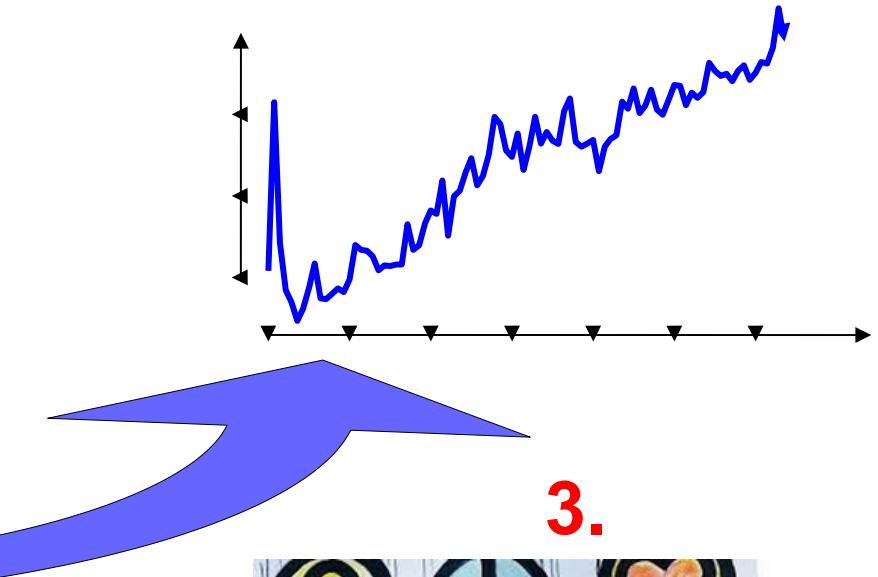
information diffusion
epidemics spreading
behavioral modeling

2.



[Rizoiu et al ICWSM'18]

[Kim et al Journ.Comp.SocSci'19]



3.



Slipping to the Extreme: A Mixed Method to Explain How Extreme Opinions Infiltrate Online Discussions

Quyu Kong,^{1,2} Emily Booth,² Francesco Bailo,² Amelia Johns,² Marian-Andrei Rizoiu^{1,2}

¹ Australian National University
² University of Technology Sydney
quyu.kong@anu.edu.au, emily.booth@uts.edu.au, amelia.johns@uts.edu.au,
marijan-andrei.rizoiu@uts.edu.au

Abstract

Qualitative research provides methodological guidelines for observing and studying communities and cultures on online social media platforms. However, such methods demand considerable manual effort from researchers and may be overly focused and narrowed to certain online groups. In this work, we propose a complete solution to accelerate qualitative analysis of problematic online speech — with a specific focus on opinions emerging from online communities — by leveraging machine learning algorithms. First, we employ qualitative methods of deep observation for understanding problematic online speech. This initial qualitative study constructs an ontology of problematic speech, which contains social media postings annotated with their underlying opinions. The qualitative study also dynamically constructs the set of opinions, simultaneous with labeling the postings. Next, we collect a large dataset from three online social media platforms (Facebook, Twitter and YouTube) using keywords. Finally, we introduce an iterative data exploration procedure to augment the dataset. It alternates between a data sampler, which balances exploration and exploitation of unlabeled data, the automatic labeling of the sampled data, the manual inspection by the qualitative mapping team and, finally, the retraining of the automatic opinion classifier. We present both qualitative and quantitative results. First, we present detailed case studies of the dynamics of problematic speech in a far-right Facebook group, exemplifying its mutation from conservative to extreme. Next, we show that our method successfully learns from

and Vraga 2018) being recorded in the literature. To date, there exist three primary types of methods for addressing problematic information. The first type concentrated on large-scale monitoring of social media datasets to detect inauthentic accounts (bots and trolls) (Ram, Kong, and Rizoiu 2021) and coordinated disinformation campaigns (Rizoiu et al. 2018). The second group aims to understand which platforms, users, and networks contribute to the “infodemic” (Smith and Graham 2019; Bruns, Harrington, and Hurcombe 2020; Colley and Moore 2020). The third group uses computational modeling to predict future pathways and how the information will spread (Molina et al. 2019). These studies provide valuable insights into understanding how problematic information spreads and detecting which sources are reshared frequently and by which accounts. Though the first and third research approaches offer breadth of knowledge and understanding, there are limitations — they often have less to say about why certain opinions and views gain traction with vulnerable groups and online communities.

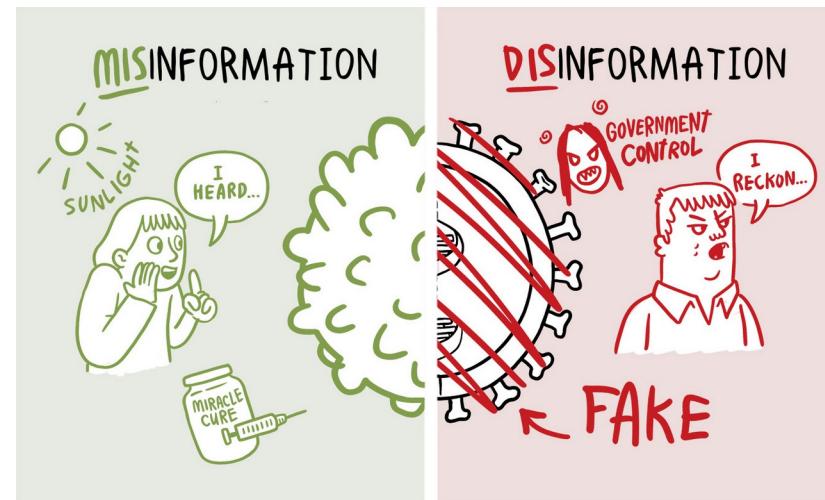
Qualitative research methods are well placed to address this gap. They provide rich, contextual insights into the social beliefs, values, and practices of online communities, which shape how information is shared and how opinions are formed (Glaeser and Sunstein 2009; Boyd 2010; Baym 2013; Johns 2020). This is also fundamental to un-



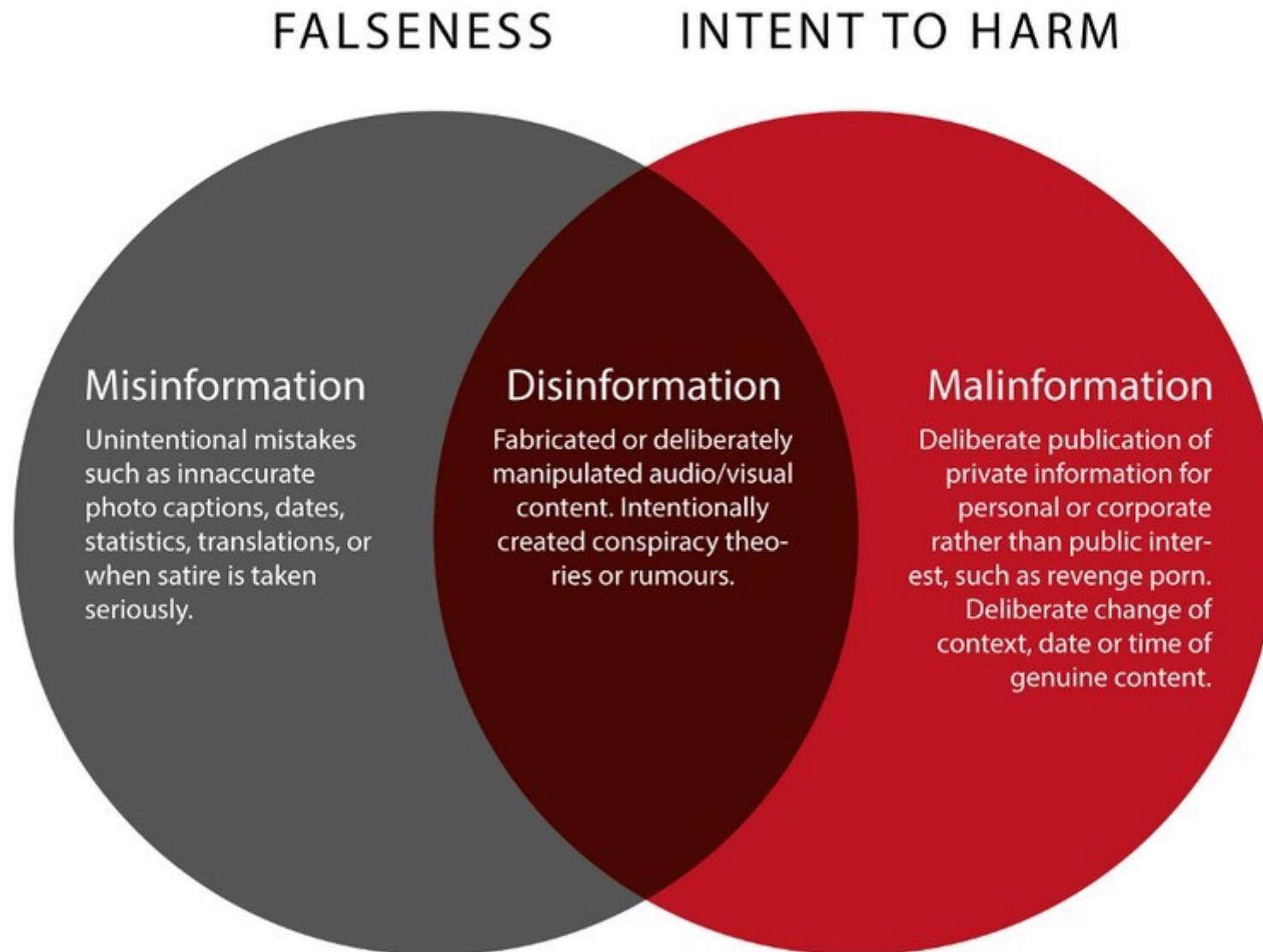
Motivation

Problematic speech is online interactions, speech, and artefacts that are inaccurate, misleading, inappropriately attributed, or altogether fabricated (Jack 2017).

- misinformation
- disinformation
- hate speech



Types of information disorder



The gap of methods



Computational and quantitative

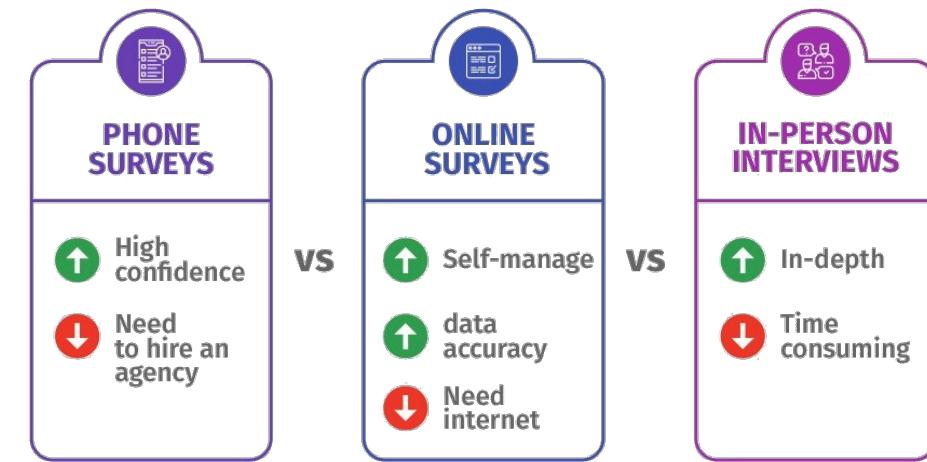
Large-scale monitoring of social media datasets

[Kong et al, CIKM'20]
[Ram et al, WSDM'21]

Identify platforms, users, and networks that contribute to the “infodemic”

[Smith and Graham 2019]
[Bruns et al 2020]

Future information spread [Molina et al. 2019]



Qualitative and ethnographic

How information is shared and how opinions are formed

[Boyd 2010] [Baym 2015]

Why opinions and information sources scale to encompass large segments of the online society

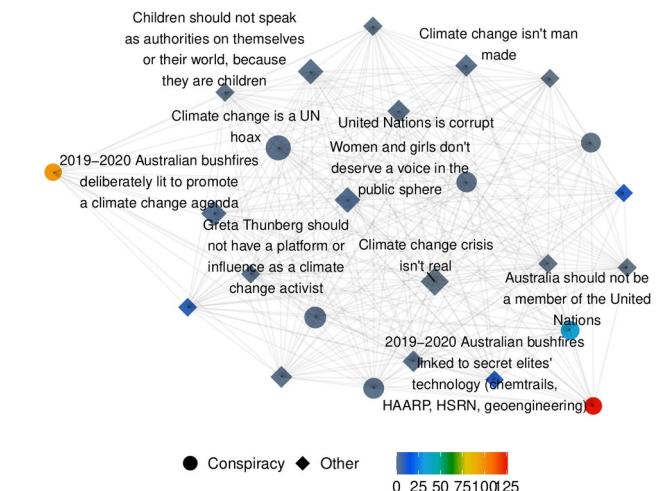
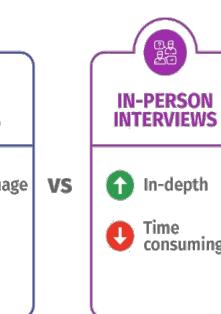
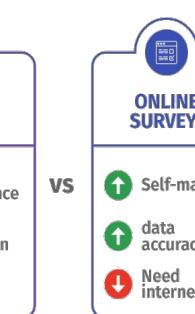
[Bailo 2020]
[Bruns et al 2020]

Research questions

Can we leverage both qualitative and quantitative analysis for studying problematic online speech?

Can we accelerate qualitative research and observations of online behavior with machine learning algorithms?

Can we track the dynamics of problematic opinions from online discussions using unlabeled data?



Interdisciplinary approach and team



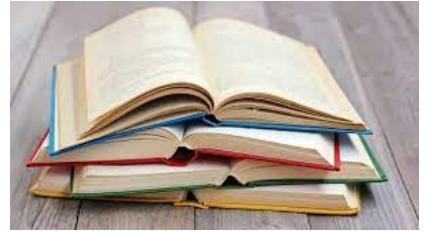
Communication science



Francesco Bailo



Amelia Johns



Literature



Emily Booth



Computer science

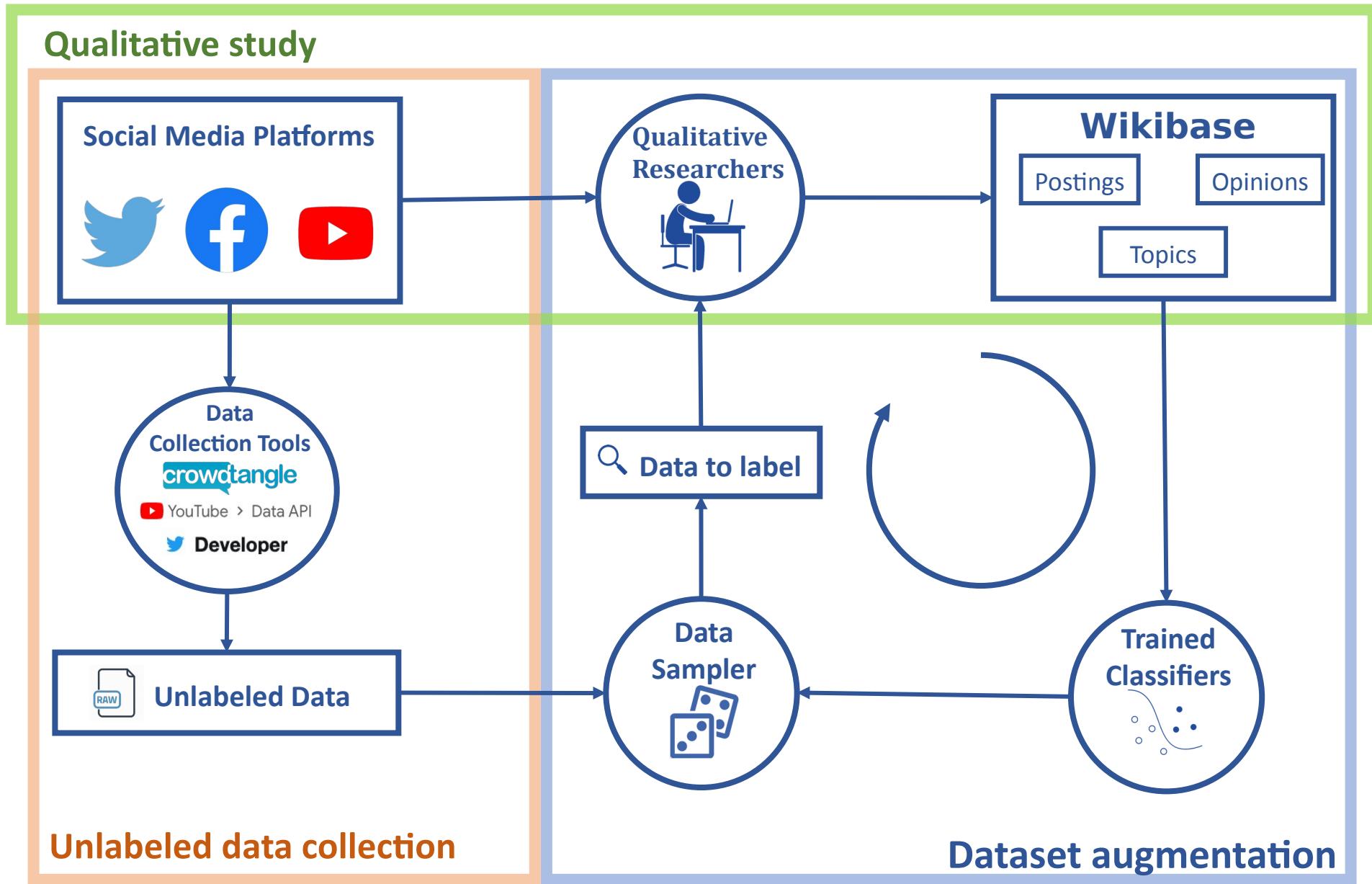


Marian-Andrei Rizoiu



Quyu Kong

Presentation Plan / Overall Approach





1. Qualitative Study



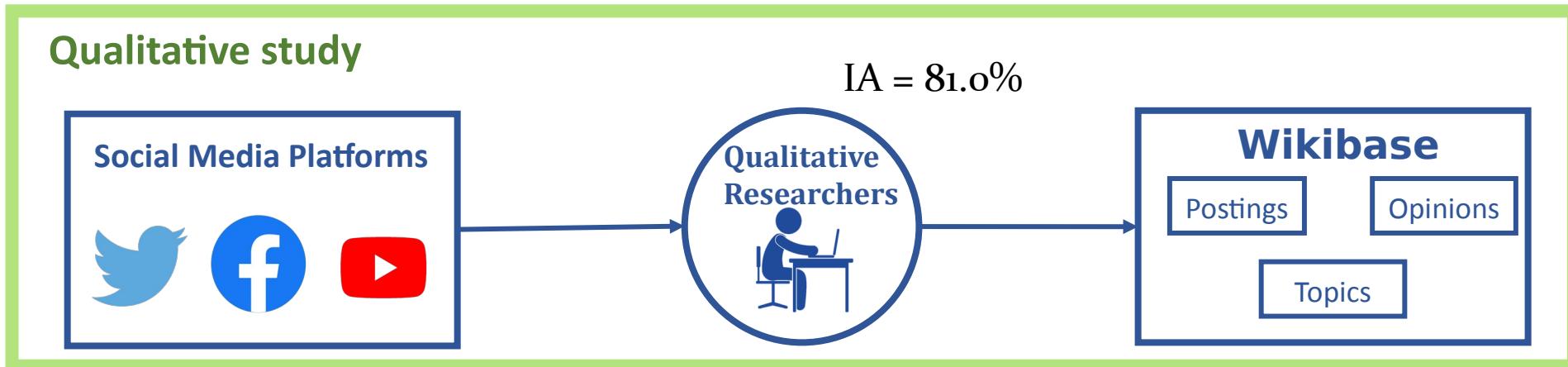
Four topics:

- 2019-20 Australian bushfire season
- Climate change
- COVID-19
- Vaccination

Dec 2019 – Jan 2021

Internet places:

- News stories
- Facebook page monitoring
- Cross-page link tracking
- Platform recommender systems



614 postings and 65 opinions:

- Climate change crisis isn't real
- United Nations is corrupt
- Climate change is a UN hoax
- United Nations want to be the global ruling government
- Experts manipulate data for private or corporate agendas
- Vaccines cause Autism
- The World Health Organization is corrupt
- Men are being chemically emasculated by the government/science/elites
- Covid-19 is the Chinese government's bioweapon

Take Australia Back - Public Facebook group, 11.2K members.

Sample post and comments 1: Jan 10 2020

A screenshot of a Facebook post and its comments from the 'Take Australia Back' group. The post, made by a user on January 10, states: "Apparently climate change is real 🤦‍♂️ Apparently half of this group are smart enough to disprove the fact tho 🤦‍♂️". It has 42 reactions and 220 comments. One comment from a user named 'David' (@#reaserchgeoengineering) asks about glaciers and sea level rise. Another user (@user123) responds, "Why im a climate change sceptic. Carbon is 3% of our atmosphere. And 0.4% of that 3% is man made. So yeah, not buying it. Especially considering some of these experts are lying. Like Sir David Attenborough lying about the walruses jumping off the cliff... See More". This comment has 5 reactions and 12 replies.

- 50/50 climate change denial and support
- Some respectful debate but mainly polarising contest and troll-like social practices
- Use of misogynistic and ableist abuse to inflame/polarise/derail opposing opinion
- **Small number of conspiracy theories (e.g. chemtrails)**
- 40-60+ user group
- **Text based comments, few links out, more comments than shares**

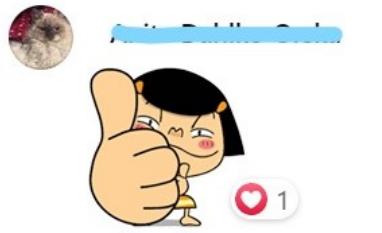
Sample post and comments 2: 16 September 2020



84 9 Comments 58 Shares

Like

Share



Like · 5h

Yeah, you can just remove the cloth masks anytime you want. And also they don't silence you as much as muffle your voice.

Like · 5h · Edited

But it covers so much of your emotion and power. There is a reason men do this to women in Islam.

Like · 5h

I'M SELFISH?

You force others to inject themselves with dangerous substances so YOU feel safe.

You force others to cover their source of oxygen for months on end so YOU feel safe.

You force others to lose their jobs & retirements so YOU feel safe.

You force others stay home so YOU feel safe.

I haven't asked one person to do one thing. YOUR list is LONG and endless.

Like · 1h

how true

I joined this group when we were fighting against scomo cause he's a dick head, now this group is full of dickheads

Like · 1h · Edited

3

Not too mention most of those iron masks had funk locks on them

Like · 1h

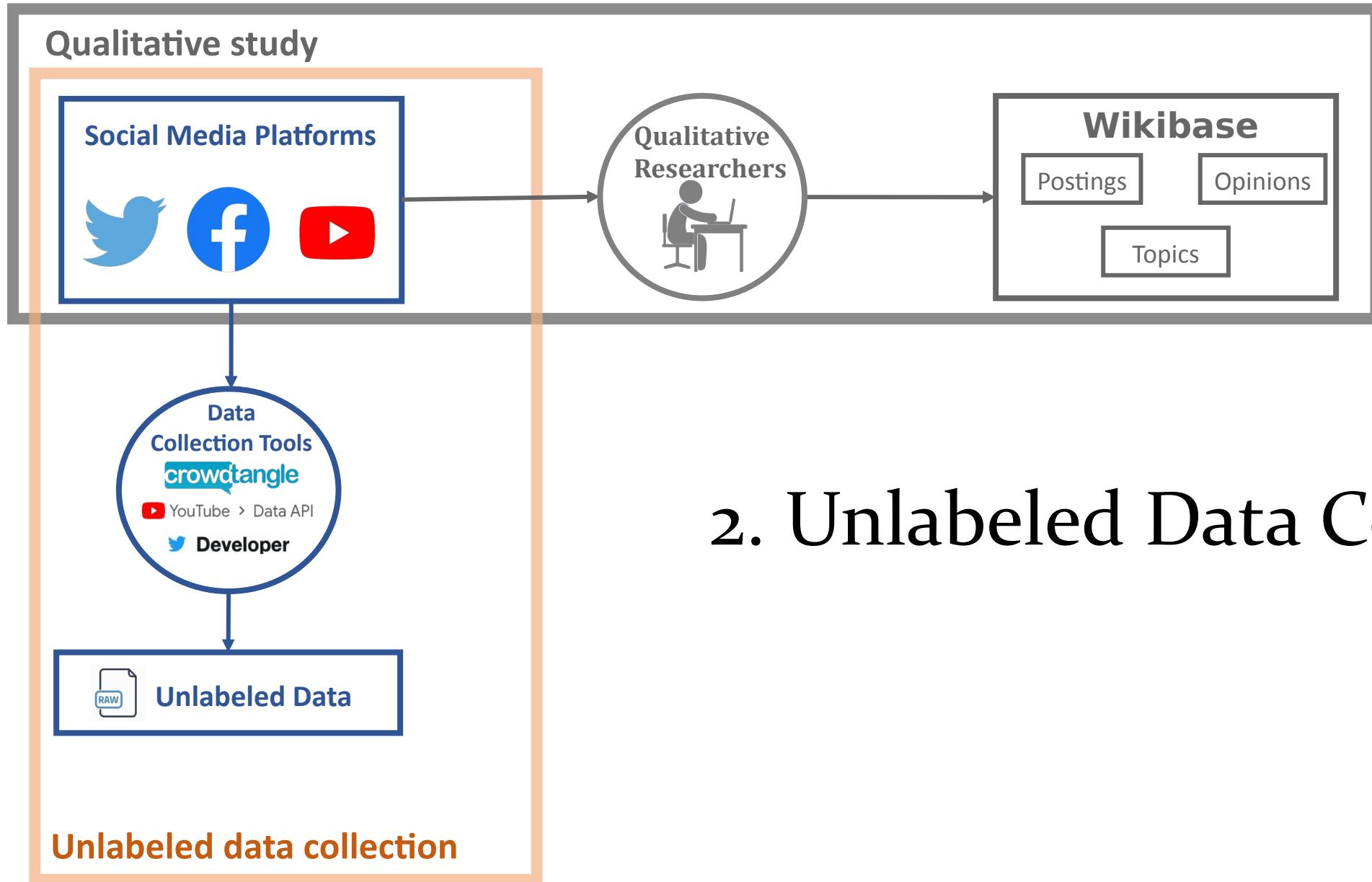
3

its one of the first thing you learn on your road of indoctrination you mean. 😊

Like · 1h

realise who controls the puppets first.





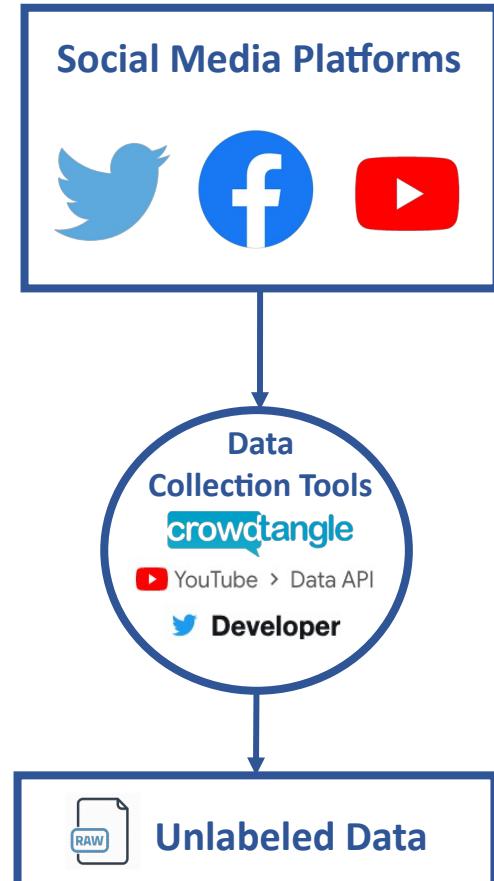
2. Unlabeled Data Collection

2. Unlabeled Data Collection

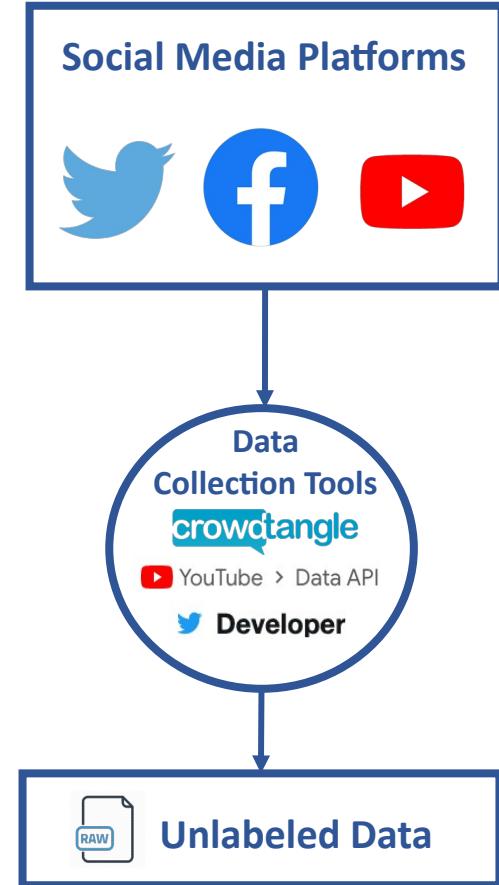
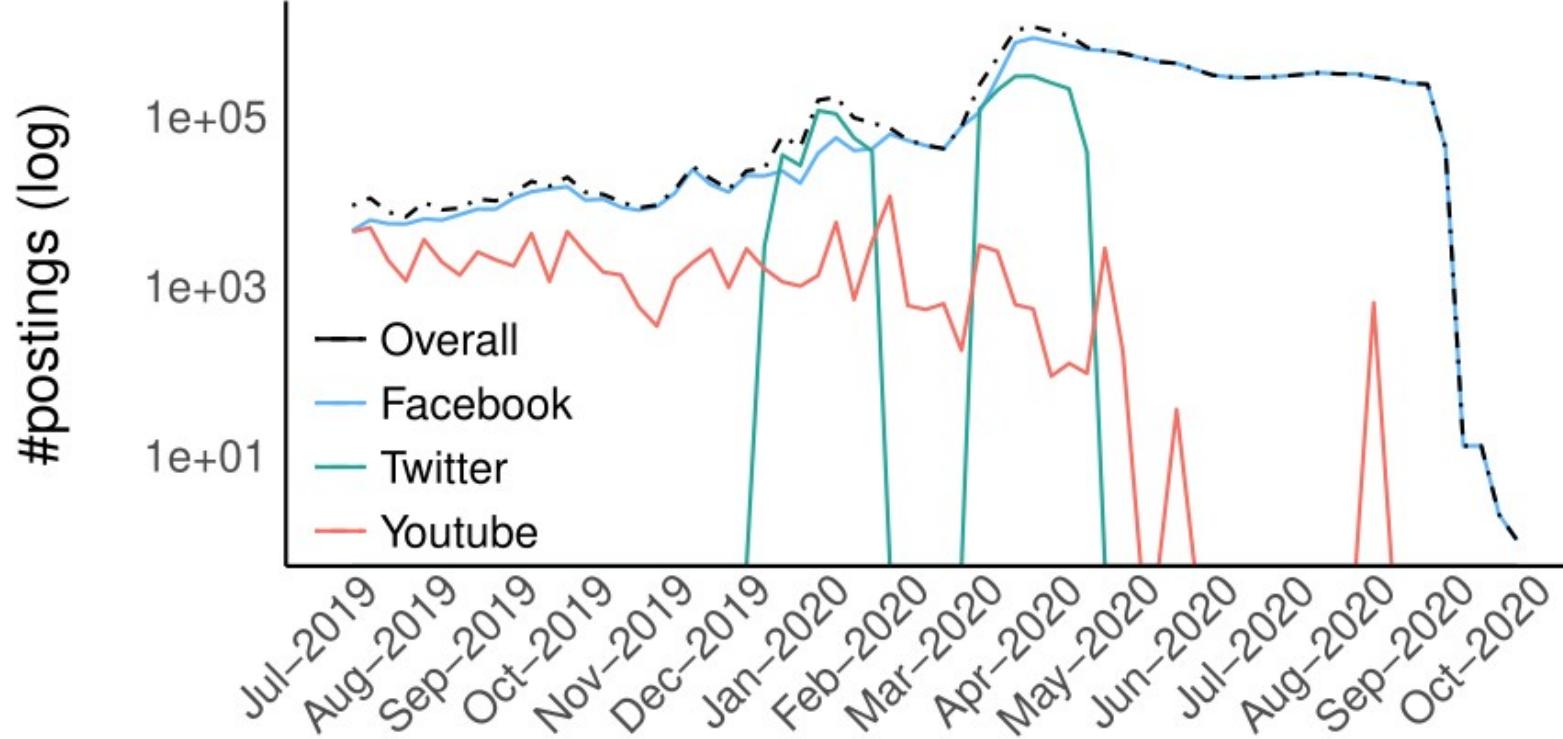
Topics	Selected keywords
2019-20 Australian bushfire season, Climate change	bushfire, australian fires, arson, scottymarketing, liarfromtheshiar, australiaburns, australiaburning, itsthegreensfault, backburning, back burning, climate change, climate emergency, climate hoax, climate crisis, climate action now
Covid-19, Vaccination	covid, coronavirus, covid-19, pandemic, world health organization, vaccine, social distancing, quarantine, plandemic, chinavirus, wuhan, stayhome, MadeinChina, ChinaLiedPeopleDied, 5G, chinacentric

13.3M postings:

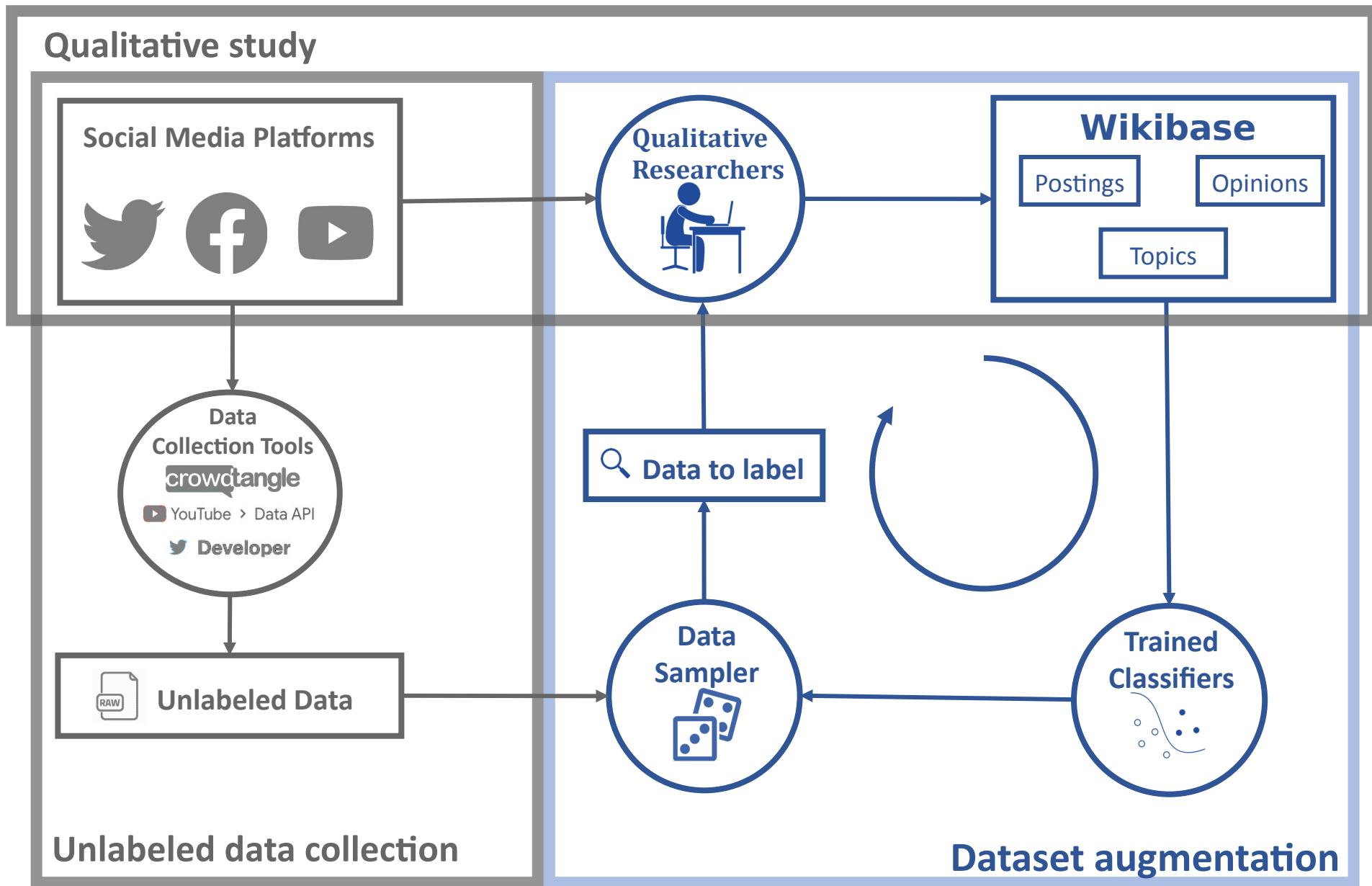
- 11.4M Facebook
- 1.8M Twitter
- 91K YouTube comments



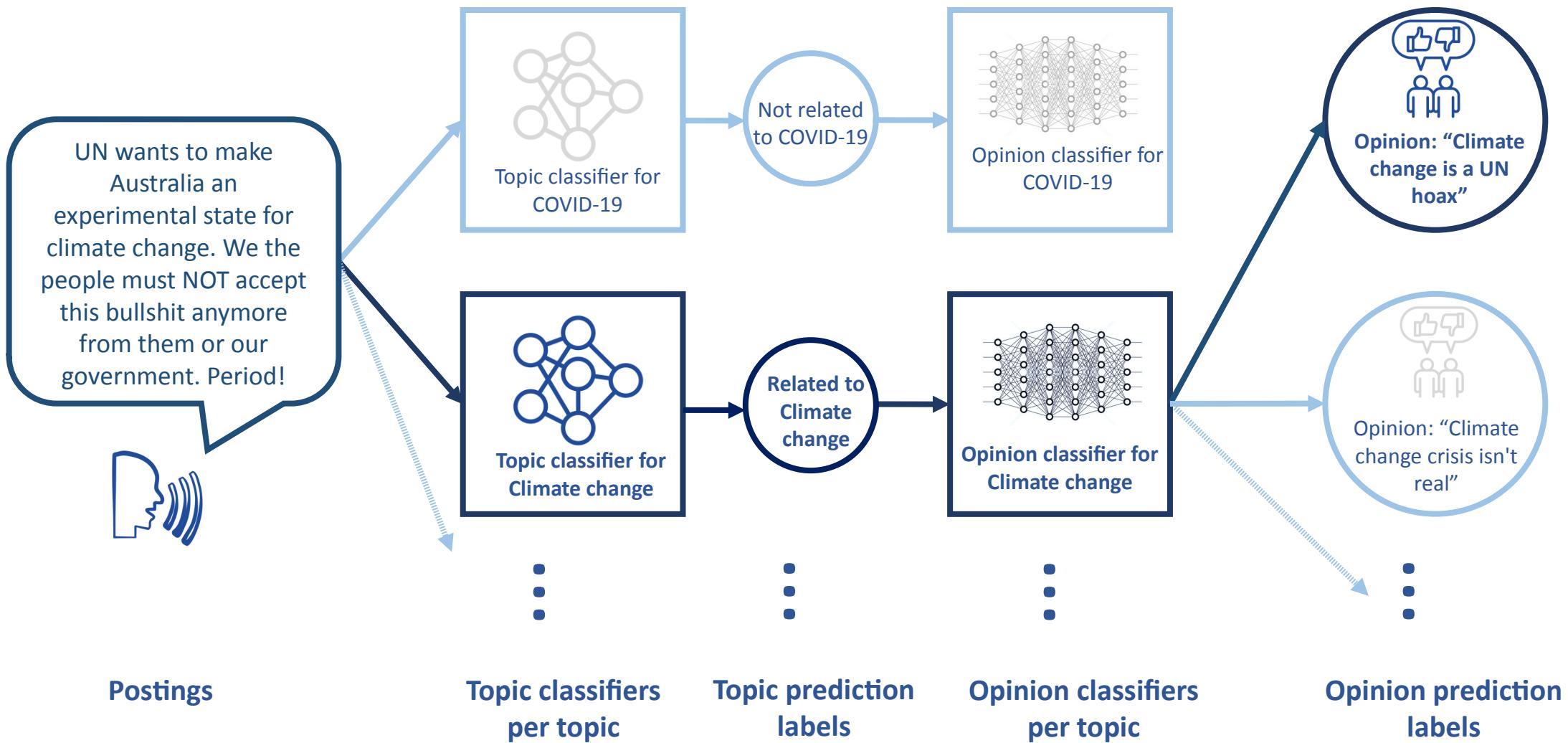
2. Unlabeled Data Collection



3. Dataset Augmentation



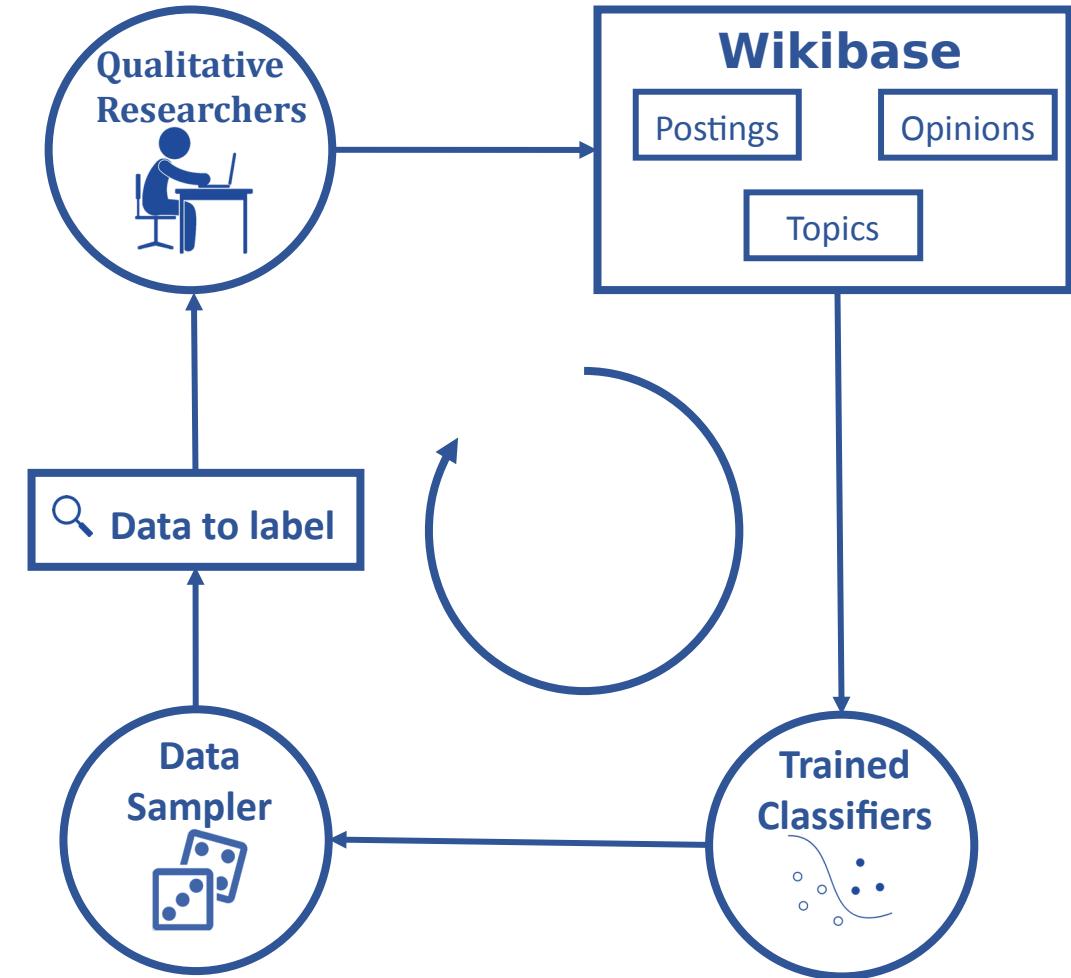
Two levels of classifiers



	RF	SVM	XGBoost	RoBERTa
Macro Accuracy	0.791	0.775	0.779	0.800
Macro F1	0.782	0.768	0.768	0.800

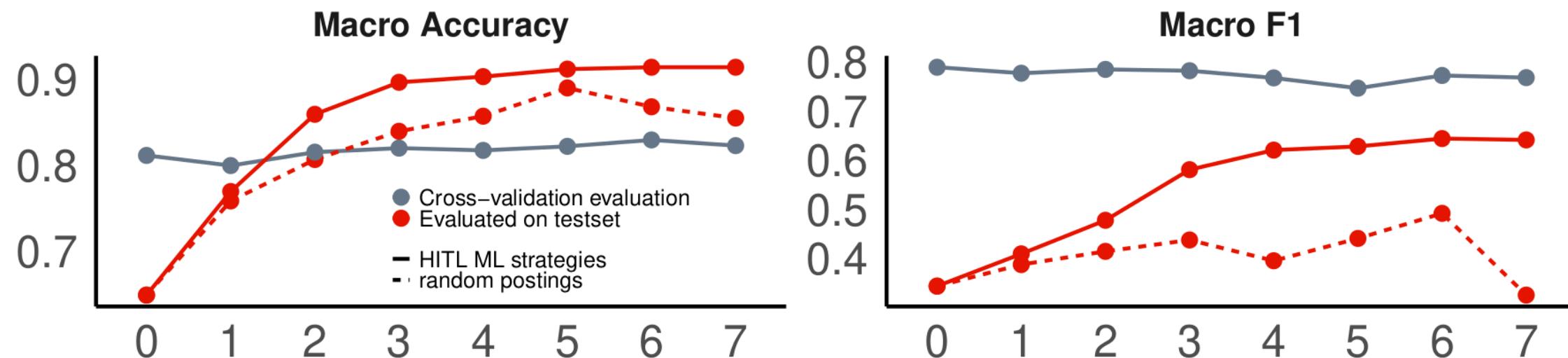
3. Dataset Augmentation

- Human-in-the-loop Machine Learning
- Three strategies for data sampling:
 - Active learning
10 posts / iteration / topic
 $u(\mathbf{x}) = 1 - p(\hat{y} \mid \mathbf{x}; f_{t,i})$
 - Top confidence
10 posts / iteration / topic
 - Random sampling
5 posts / iteration / topic
- Iterated until convergence
 - cross-validation error VS test set error
 - gain on test set between two iterations



Results

Human-in-the-loop performance



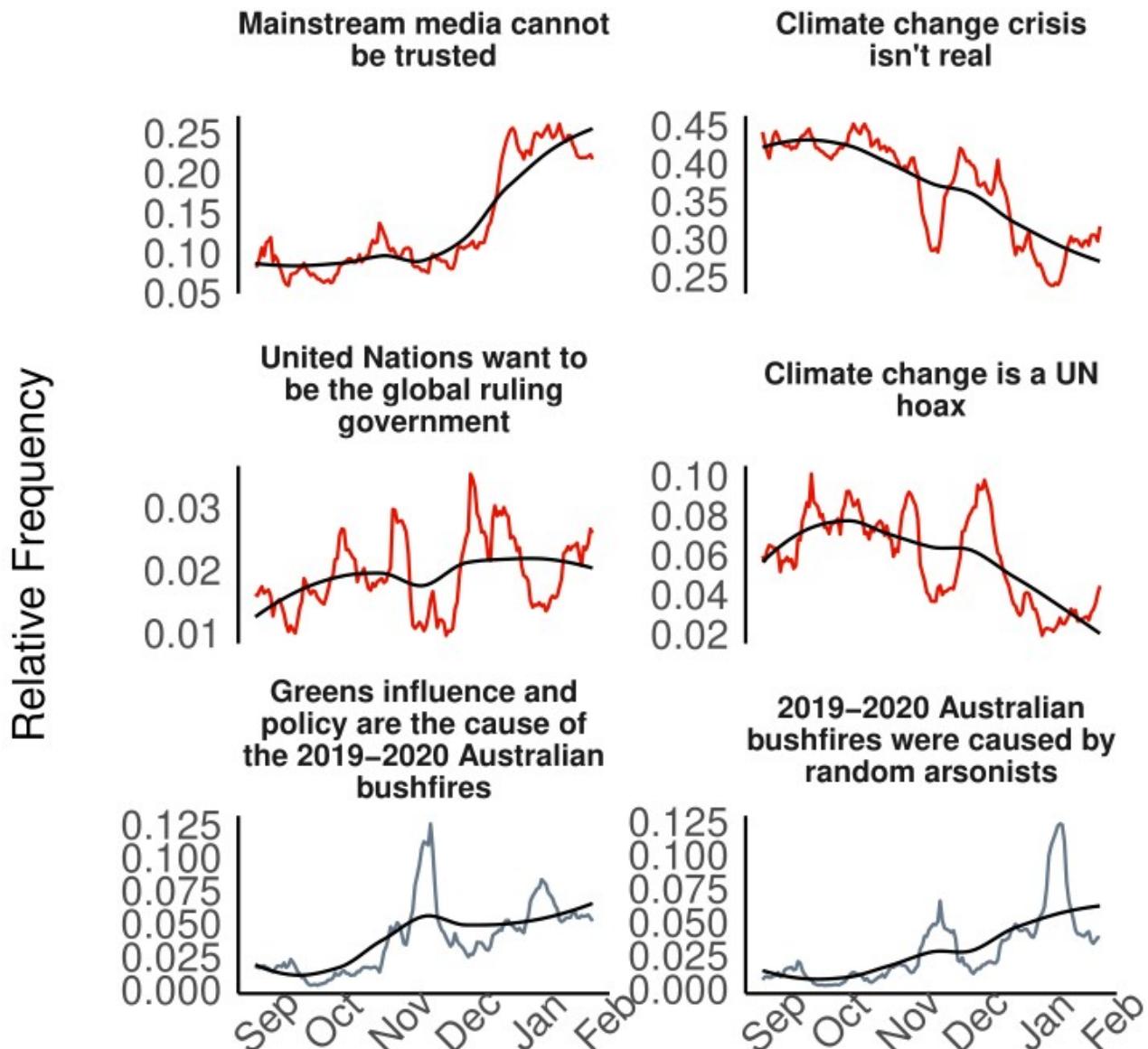
- Performances improve as more batches are performed
- Gap between generalization and test set error reduces
- Improvement plateaus as the process converges
- Human-in-the-Loop outperforms static random selection of samples

	L0	L7
#posts	614	1381
#opinions	65	71

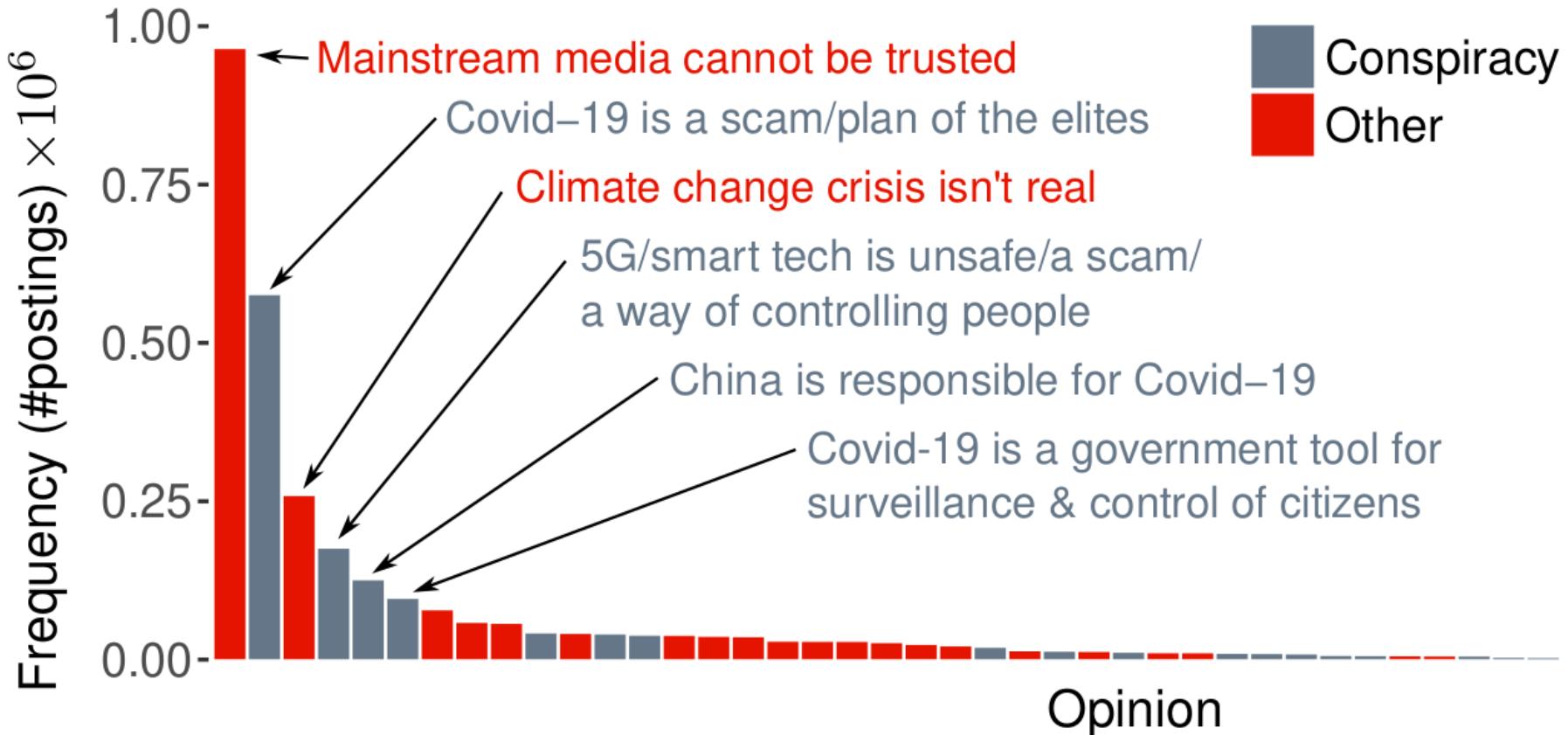
Opinion analysis at scale

Fully labeled dataset stats

- 1.7M postings with at least one opinion
- 314K postings with 2 or more opinions
- 21.26M off-topic postings
- **Total: 22.96M postings**



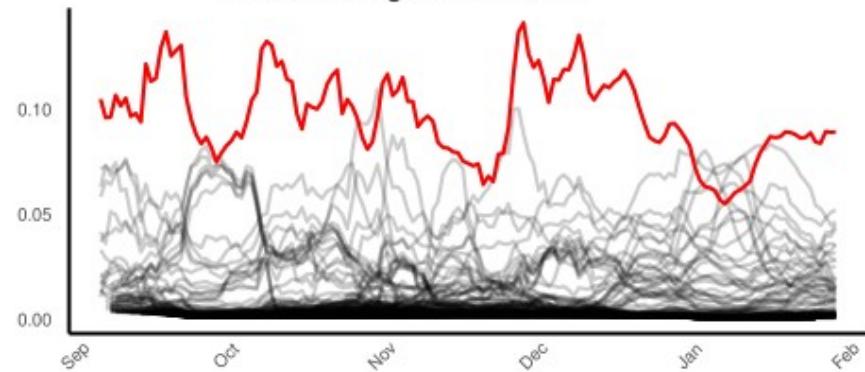
Opinion analysis at scale



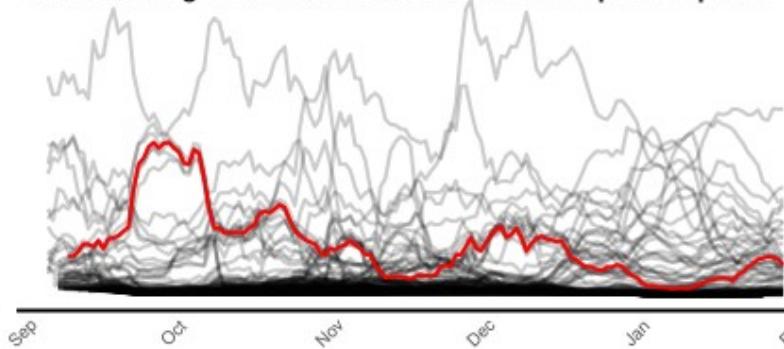
- Opinion usage frequency is longtail distributed
 - Four of the top six opinions endorse conspiracy theories

Opinion co-occurrence network

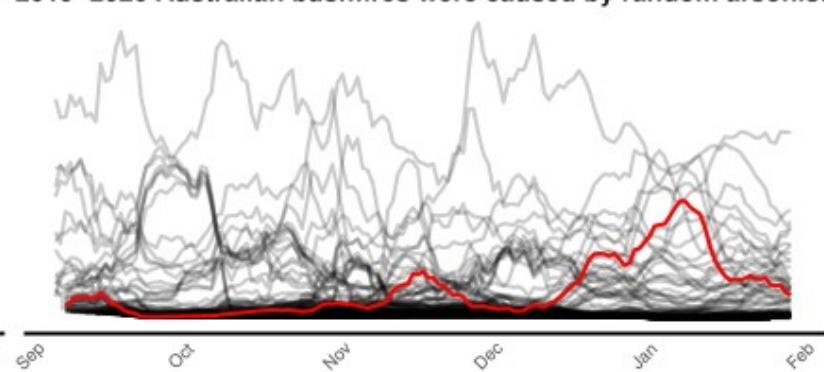
Climate change crisis isn't real
Climate change is a UN hoax



Greta Thunberg should not have a platform or influence as a climate change activist
Women and girls don't deserve a voice in the public sphere



2019–2020 Australian bushfires and climate change not related
2019–2020 Australian bushfires were caused by random arsonists



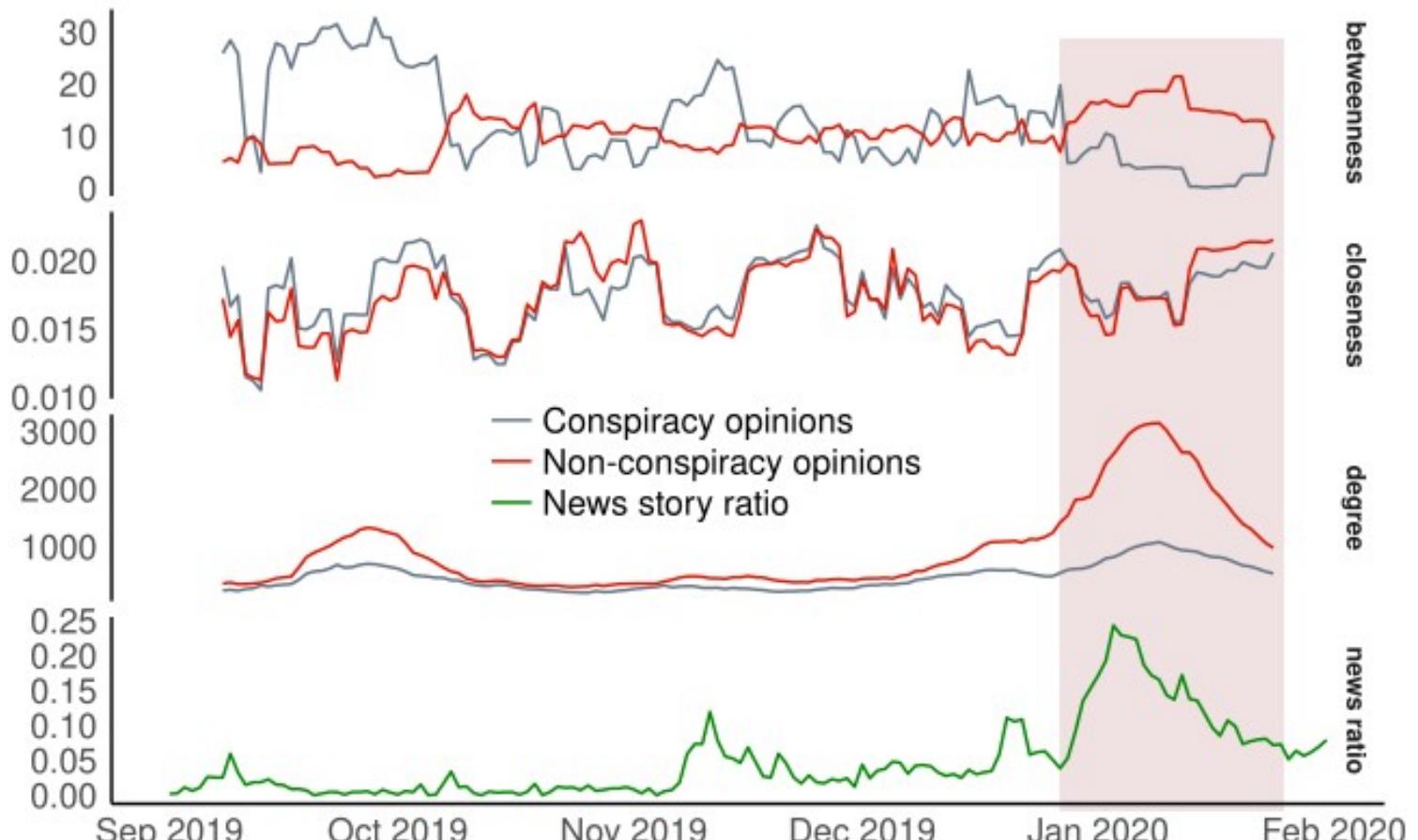
A continuous and relatively strong association between prevalent opinions

Associations with declining relative frequencies

Rising associations – early warnings for their adoption (and possibly normalization) by participants

Centrality of conspiracy opinions and news ratio

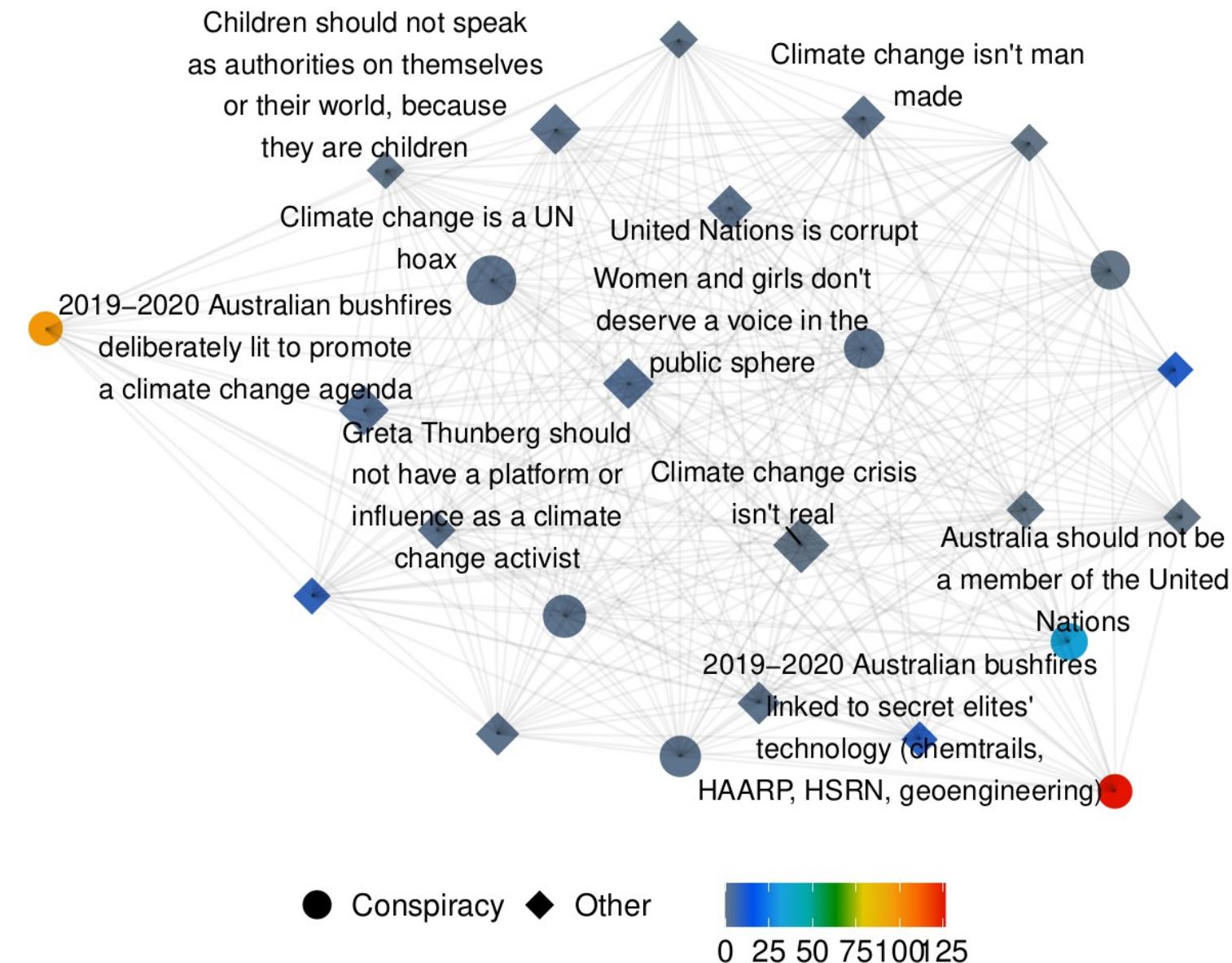
Higher coverage from news media reduces centrality of conspiracy opinions.



coverage ratios from Media Cloud
(Roberts et al. 2021)

Opinion co-occurrence network

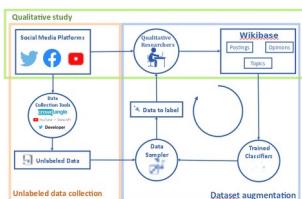
- High betweenness centrality of conspiracy opinions → selectively used in conjunction with many other opinions
- 14 days in late September 2019 – peak betweenness
- Conspiracy opinions are used together with mainstream opinions – rationalize and popularize them



Summary



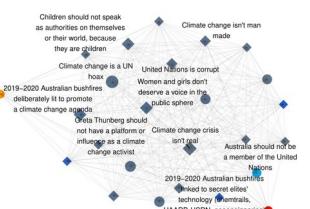
An inter-disciplinary team and methods to solve a difficult task: detecting and mapping the impact of online problematic content



A mixed qualitative and human-in-the-loop Machine Learning approach for detecting problematic content



A representative annotated dataset of online problematic content, qualitative and quantitative analyses



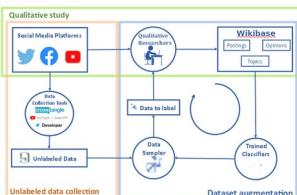
A hypothesis of how fringe opinions infiltrate mainstream discourse via co-occurrence with established opinions



Thank you!



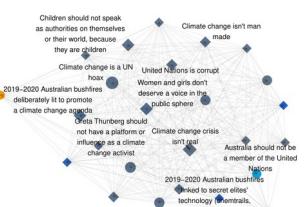
An inter-disciplinary team and methods to solve a difficult task: detecting and mapping the impact of online problematic content



A mixed qualitative and human-in-the-loop Machine Learning approach for detecting problematic content



A representative annotated dataset of online problematic content, qualitative and quantitative analyses



A hypothesis of how fringe opinions infiltrate mainstream discourse via co-occurrence with established opinions