

# Computational methods for querying and sampling the Twitter disinformation datasets

Partnership for Countering Influence Operations Workshop, 201124  
[nils.holmberg@isk.lu.se](mailto:nils.holmberg@isk.lu.se)

# Background

- media effects, psychology
- identify, counter disinfo
- where is disinfo prevalent?
- where can we access content?
- social media?
- not easy to find, define
- Twitter datasets (James)



## Fakta om naturen

Träd växer **överallt förutom där** det permanent är is och snö, på topparna av höga berg samt i öknen. Lämnas en bit mark orörd tillräckligt länge kommer det där med tiden att börja växa upp träd. Till att börja med täcks jorden av lågt växande krypväxter. Efterhand växer buskar fram, dessa kommer då att döda delar av de låga växterna eftersom de hamnar i skuggan. Efter ytterligare någon tid börjar träden växa upp och dessa tar i sin tur död på några av buskarna eftersom de hamnat i skuggan från bladverket. På detta vis uppstår nya skogar med tiden. Många träd växer väldigt långsamt och några har en väldigt lång livslängd. När gamla träd dör växer det fram yngre träd i dess ställen. Skogar är levande platser som kan förbli oförändrade under långa perioder, vilka träd som kan hittas i en skog beror på omgivningens klimat.

[Klicka här när du är klar!](#)



# Querying datasets

- data aggregation, summary
- requires data normalization
- unique id, one record / row
- solution: json format, sql
- aggregate by language, content
- possibility, brand hijacking?
- problems: user count mismatch

```
"tweet_id"    "tweet_time"    "tweet_text"
"908898221715447808"    "2017-09-16 03:39"    "The
Myanmar military is deliberately burning ethnic
#Rohingya villages near the
#Bangladesh border, HRW said on Friday.
#RohingyaCrisis"
```

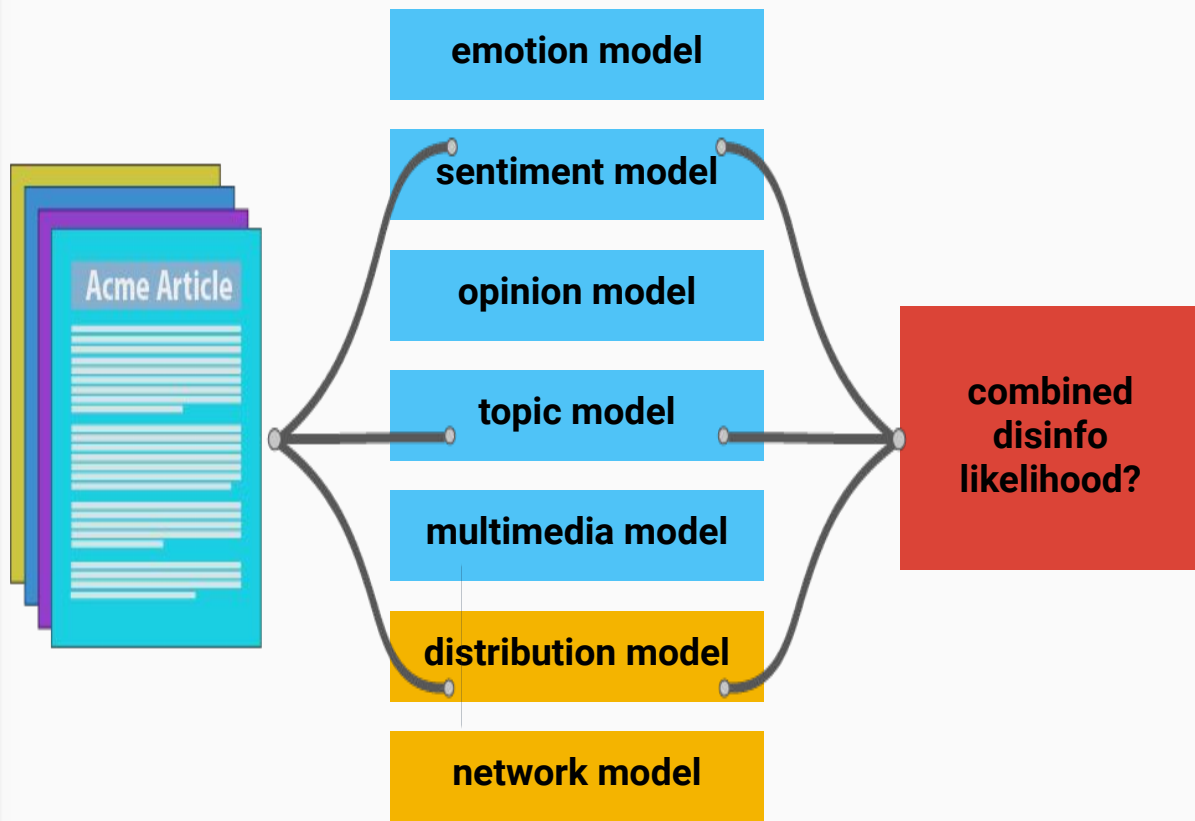
```
{
  "tweet_id": "908898221715447808",
  "tweet_time": "2017-09-16 03:39",
  "tweet_text": "The Myanmar military is deliberatel
y burning ethnic \n#Rohingya villages near the \n#
Bangladesh border, HRW said on Friday. #RohingyaCris
is"
}
```

The aggregated data in response to RQ3 (Table 1) shows tweet counts by region, user counts, average proportion of English tweets, and average number of tweets per user (N=210M).

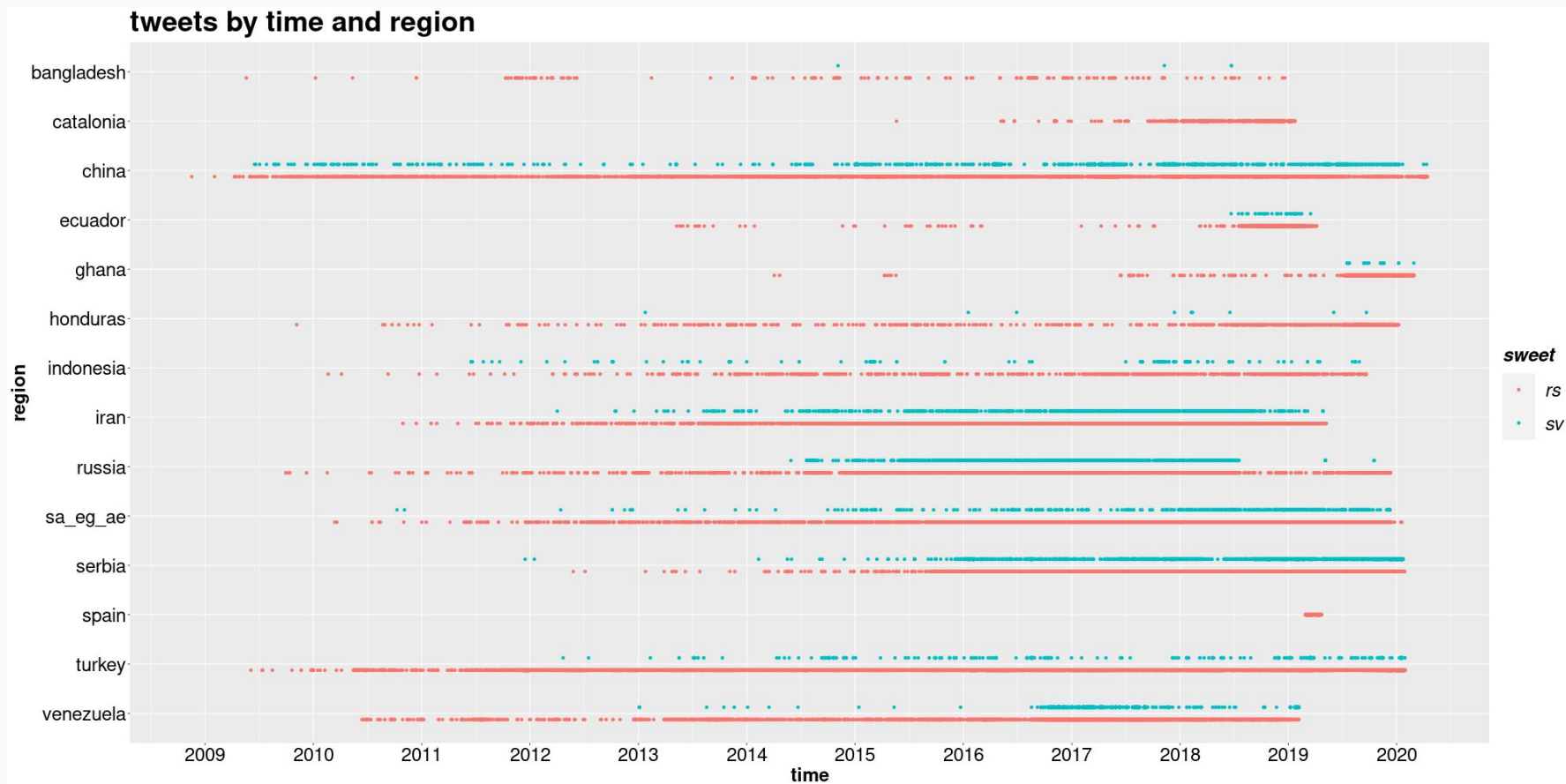
region.....	t_count..	t_count_en.	t_count_sv.	u_count..	.t_prop_en	.t_per_u
:-----	:-----	:-----	:-----	:-----	-----:	-----:
bangladesh.	26.21.K..	26.21.K....	4.....	11.....	.....1.00	.2382.91
catalonia..	9.49.K..	2.04.K....	0.....	76.....	.....0.22	..124.86
china.....	14.2.M...	7.42.M....	2.56.K....	29.64.K..	.....0.52	..478.97
ecuador....	700.24.K.	89.K.....	91.....	787.....	.....0.13	..889.76
ghana.....	39.96.K..	39.96.K....	22.....	60.....	.....1.00	..666.07
honduras...	1.17.M...	290.61.K..	32.....	3.01.K..	.....0.25	..387.44
indonesia..	2.7.M...	1.6.M....	1.6.K....	716.....	.....0.59	.3771.36
iran.....	9.31.M...	7.01.M....	4.63.K....	9.83.K..	.....0.75	..947.88
russia.....	13.12.M..	5.76.M....	4.75.K....	4.86.K..	.....0.44	.2702.67
sa_eg_ae...	78.05.M..	18.06.M....	6.32.K....	35.7.K..	.....0.23	.2186.64
serbia.....	43.07.M..	18.83.M....	4.22.K....	42.16.K..	.....0.44	.1021.51
spain.....	56.71.K..	20.64.K....	0.....	216.....	.....0.36	..262.56
turkey.....	36.95.M..	3.04.M....	4.49.K....	22.52.K..	.....0.08	.1640.70
venezuela..	10.5.M...	1.66.M....	1.58.K....	2.48.K..	.....0.16	.4244.44
<b>total.....</b>	<b>209.91.M.</b>	<b>63.85.M....</b>	<b>30.3.K....</b>	<b>152.05.K.</b>	<b>.....0.44</b>	<b>.1550.56</b>

# Sampling datasets

- 1000 tweets per archive
- analyze content, activity
- understand disinfo signals
- **content**
- sentiment, emotion
- topic modelling, longer docs?
- word2vec approach
- obj. detection (Zannettou 2019)
- **activity**
- social network analysis
- frequency over time graphs
- tweet time relative to events
- interface for selecting data?



Tweets by time and region in response to RQ2 (Figure 1). Red dots indicate English language tweets, while green dots indicate tweets explicitly referring to Sweden (N=60K sample).



# Sampling datasets

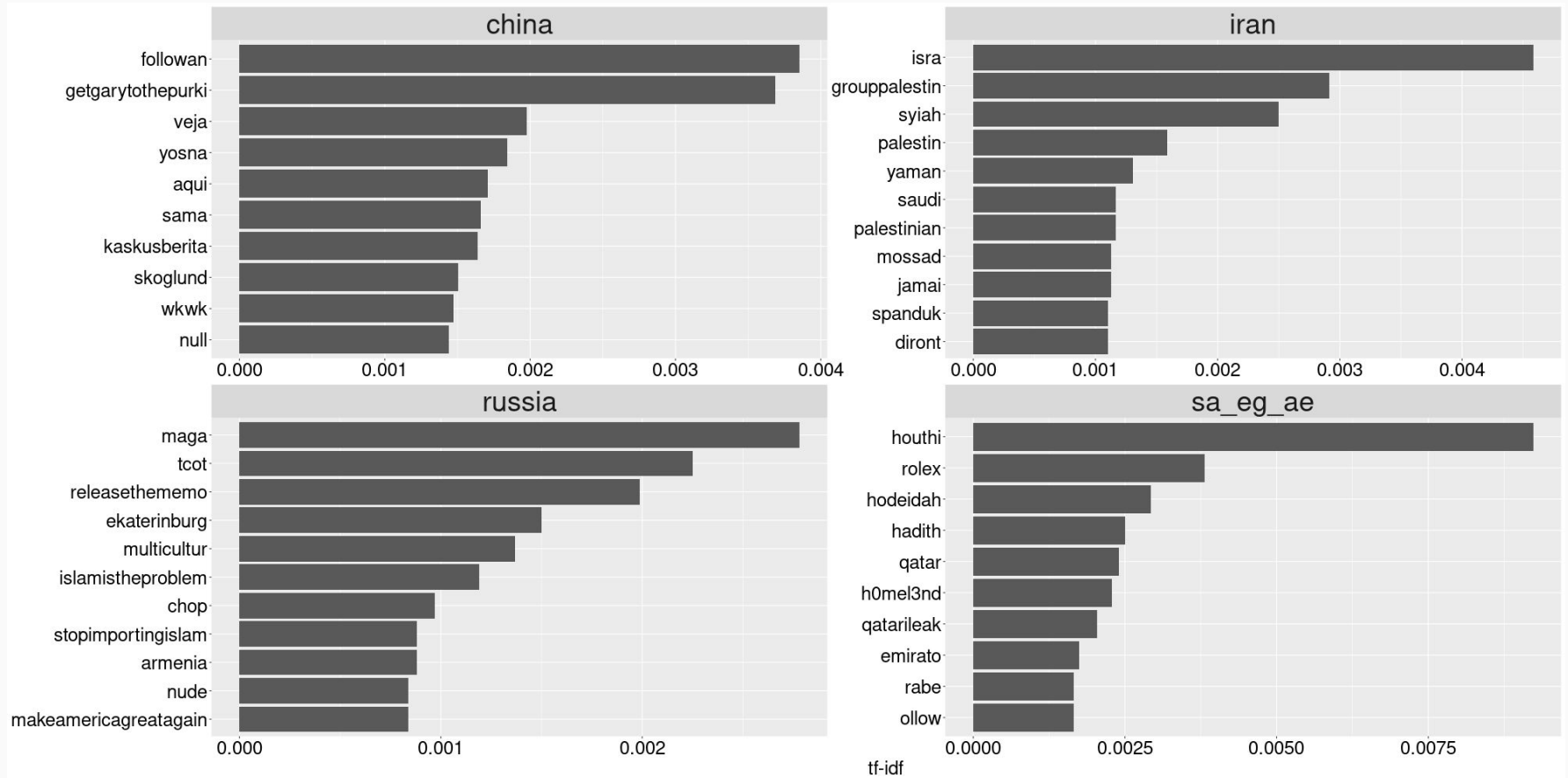
- 1000 tweets per archive
- analyze content, activity
- understand disinfo signals

- **text content**
- important words, tf-idf
- word weights
- used for word clouds

- **predict region**
- china, iran, russia, sa\_eg\_ae
- english text content?

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	$\log(2/2) = 0$	0	0
Car	1/7	0	$\log(2/1) = 0.3$	0.043	0
Truck	0	1/7	$\log(2/1) = 0.3$	0	0.043
Is	1/7	1/7	$\log(2/2) = 0$	0	0
Driven	1/7	1/7	$\log(2/2) = 0$	0	0
On	1/7	1/7	$\log(2/2) = 0$	0	0
The	1/7	1/7	$\log(2/2) = 0$	0	0
Road	1/7	0	$\log(2/1) = 0.3$	0.043	0
Highway	0	1/7	$\log(2/1) = 0.3$	0	0.043

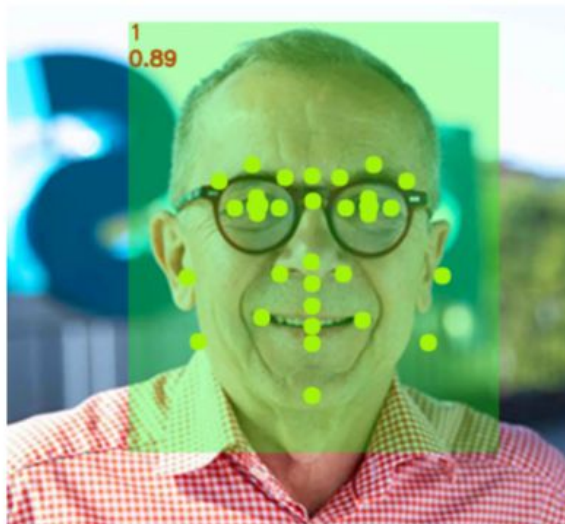
The most important words in tweet text according to TF-IDF weighting in response to RQ4 (Figure 2). Comparison between four prolific disinfo regions (China, Iran, Russia, SA\_EG\_AE, N=60K sample).





# Future research

- currently, preprocessing
- disinfo vs [internet archive](#)
- hand labelling data
- countering, media effects
- detection
- understanding
- countering




1 angerLikelihood VERY\_UNLIKELY  
1 blurredLikelihood VERY\_UNLIKELY  
1 headwearLikelihood VERY\_UNLIKELY  
1 joyLikelihood LIKELY  
1 sorrowLikelihood VERY\_UNLIKELY  
1 surpriseLikelihood VERY\_UNLIKELY  
1 underExposedLikelihood  
VERY\_UNLIKELY



0 angerLikelihood VERY\_UNLIKELY  
0 blurredLikelihood VERY\_UNLIKELY  
0 headwearLikelihood VERY\_UNLIKELY  
0 joyLikelihood VERY\_LIKELY  
0 sorrowLikelihood VERY\_UNLIKELY  
0 surpriseLikelihood VERY\_UNLIKELY  
0 underExposedLikelihood  
VERY\_UNLIKELY

thanks!





Our principal goal is to bridge gaps between researchers and platforms, so it would be great to focus on your most tangible findings and conclusions.

Methodology can of course be part of the story (especially if yours is particularly innovative or contentious), hopefully not the bulk.

Finally, if you can include any direct policy implications, such as how Twitter could operationalize your research, that would be most welcome—though we understand that not all research lends itself easily to that.