Epitweetr: Early warning tool using Twitter data

Laura Espinosa, Ariana Wijermans, Thomas Mollet

ECDC consultation on digital technologies, 25 May 2021

Outline

- Why epitweetr?
- Epitweetr project timeline
- Epitweetr characteristics
- Epitweetr community

Epidemic intelligence at ECDC

> **Epidemic intelligence (EI)** is the early identification of potential health threats, as well as their verification, assessment and investigation

Epidemic intelligence sources

- Indicator-based surveillance (IBS)
  - **National/regional surveillance systems** (TESSy), syndromic surveillance, lab data…
- Event-based surveillance (EBS)
  - **Official public sources** (WHO, National PH agency websites)
  - **Informal sources**
    - Web aggregators
      - Moderated (ProMED, GPHIN)
      - Automatic (MediSys, HealthMap, EIOS)
    - Social media
      - Manual monitoring (Twitter, Facebook, Flu trackers…)
      - Automated monitoring (Twitter)
  - **Restricted platforms** (EWRS, RASFF, EPIS, WHO IHR, ADSN)

Manual monitoring of social media for epidemic intelligence activities at ECDC

Manual monitoring:

ECDC is following approx. 100 Twitter and 40 Facebook accounts

ECDC has mapped the main Twitter and Facebook accounts of interest

%

30-50% of signals from social media, mainly Twitter (30% of 500 sources for COVID-19)

Event validation through Facebook or Twitter

Why automated monitoring of Twitter for epidemic intelligence?

Increased sensitivity

- Detect low signals
- Detect signals from places with low media coverage

Reduces human intervention

- Not dependent on preselected list of users
- Tool will alert us, instead of us searching

Improved timeliness

- Near real-time signal detection

Epitweetr project timeline

* https://www.medrxiv.org/content/10.1101/2021.03.25.21254342v1

epitweetr: : Early detection of public health threats from Twitter data

- **First tool** that allows to **automatically monitor trends** of Twitter data (**tweets, retweets and quotes**) by time, place and topic
- Aim to **early detect** public health threats through **signals** (unusual increase in the number of tweets)
- **Machine learning** components: geolocation, signal detection, email alerts
- **R-based, but no previous R knowledge required**
- **Free, open-source (EUPL-1.2)**
- **Legal:** ECDC Internal Procedure and Data Protection Impact Assessment

Source: https://www.ecdc.europa.eu/en/publications-data/epitweetr-tool ; https://github.com/EU-ECDC/epitweetr

How does epitweetr work?

Install epitweetr

CRAN R
The Comprehensive R Archive Network

Set up configuration

One-time downloads

Data collection

Geolocation

Aggregation

Signal detection

epitweetr interactive web application

epitweetr | Dashboard | Alerts | Geotag evaluation | Configuration | Troubleshoot

Topics

Brucellosis

Countries & regions

World (all) China

Period

custom

Dates

2020-08-07 to 2020-10-08

Time unit

Days Weeks

Include retweets/quotes

Location type

Tweet User Both

Signal false positive rate

0 0.025 0.3

0 0.03 0.06 0.09 0.12 0.15 0.18 0.21 0.24 0.27 0.3

Outlier false positive rate

0 0.05 0.3

0 0.03 0.06 0.09 0.12 0.15 0.18 0.21 0.24 0.27 0.3

Data image

Data image

Top words figure only considers tweet location, ignoring the location type parameter

Data image

Map disclaimer. Projection: +proj=robin

epitweetr community

downloads 4384

EU-ECDC / epitweetr

Unwatch 17 | Unstar 16 | Fork 3

Code | Issues | Pull requests | Discussions | Actions | Projects | Wiki | Security | ...

1. General
Welcome to epitweetr!
lauespinosa

2. Training material
Extracting Twitter data ...
lauespinosa

4. Ideas and feedback
Survey on epitweetr us...
lauespinosa

1. What is your affiliation?

National public health institution or ministry of health	28%	2
Other national institutions or ministries	14%	1
International public health organisation	42%	3
Other international organisations	0%	
Research institution	14%	1
Other	0%	

2. What is the scope of the topics you are monitoring with `epitweetr`?

Infectious diseases (communicable diseases)	60%	3
Non-communicable diseases	20%	1
Missinformation and disinformation	0%	
Drug use and addiction	0%	
Vaccines, medicines and medical devices	0%	
Sentiment analysis	0%	
Other	20%	1

3. What is your main objective for using `epitweetr`?

Learn and try epitweetr's functionalities	0%	
Assess epitweetr's usability and relevance for my work	20%	1
Use epitweetr for a specific project or task	0%	
Use epitweetr for my daily work	40%	2
Implement the use of epitweetr in my team	20%	1
Get inspired by epitweetr for developing a similar tool	20%	1

Source: https://github.com/EU-ECDC/epitweetr/discussions; https://cranlogs.r-pkg.org/badges/grand-total/epitweetr

Thank you for your attention

Additional slides

Tweet collection

Twitter Standard Search API version 1.1

- Limited to tweets in the past ~7 days
- Not exhaustive (focus on relevance)
- It is free and sufficient to meet objectives of tool

Configure topics/keywords (topics.xlsx) → in Shiny app

	A	B	C
1	#	Topic	Query
2	1	Measles	measles OR sarampion OR rougeole OR sarampo OR gafeira OR morrinha
3	2	Rubella	rubella OR rubeola OR rubeole OR rubeola OR roseola
4	3	Mumps	mumps OR parotitis OR paperas OR oreillons OR parotidite OR papeira OR caxumba
5	4	Dengue	dengue OR denv OR den-1 OR den-2 OR den-3 OR den-4 OR den-5
6	5	Haemorrhagic fever	hemorrhagic fever OR haemorrhagic fever OR vhf OR fiebre hemorragica OR fhv OR fievre

Detecting signals with epitweetr: Process overview

For each topic:
- Location
- Number of tweets
- Threshold
- Topwords

epitweetr monitors Twitter data: by topic, time & place

If number of tweets surpasses threshold: **epitweetr sends alert**

EI officer checks alert

EI officer **checks variables** in alert for EI filtering criteria

If needed, EI officer **searches** for **related tweets**

EI officer **continues EI steps** as usual (validation, descriptive analysis, assessment…)

Email or Shiny App

Twitter or R code

epitweetr and machine learning

Geolocation (tweet)

- Two steps (scores)
- Language vectors, fasttext, labelled database
- Geonames, VSM variant, lucene

Signal detection

- Modified EARS algorithm (mean, SD)
- Down-weighting outliers
- Bonferroni correction

Email alerts

- Subscribers
- Near-real-time and scheduled
- Topic and regions/countries

Preliminary results and use

Signals not detected by other sources

Ebola negative test of haemorrhagic fever – South Sudan

Signals detected earlier than others

Brucellosis outbreak from biopharmaceutical plant – Lanzhou, China

Geolocation (tweet)

~ 80-90% gets geolocated
~ 30% accuracy

Epitweetr evaluation: Geolocation

300 random tweets from 11, 16, 19 and 22 Sep 2020 (1,200 tweets)

Table 1. Summarised results from the tweet geolocation evaluation

Level		All results		>10 score	
		Country	Subnational	Country	Subnational
Total tweets		1200	1200	804	804
Extracting correct location	Correct hits	361	311	292	283
	Accuracy	30.1%	25.9%	36.3%	35.2%
Detecting tweets with extractable location	Sensitivity	72.6%	72.2%	56.6%	56.2%
	Specificity	51.6%	50.6%	69.2%	68.3%
	PPV	74.9%	74.2%	75.7%	74.9%
	NPV	48.6%	48.1%	48.6%	48.1%
	Prevalence	66.6%	66.3%	62.8%	62.7%

Source: https://www.medrxiv.org/content/10.1101/2021.03.25.21254342v1

Epitweetr evaluation: Signal detection

Signals from working days between 19 Oct and 30 Nov 2020

- 11,313 signals → 448 evaluated (334 events + 114 false signals)

- From 448 evaluated signals:
  - 63 COVID-19 related (48 events + 15 false signals)
  - 49 had only one tweet (24 events + 25 false signals)
  - PPV (general) was 3.0% and PPV (specific) was 74.6%

Source: https://www.medrxiv.org/content/10.1101/2021.03.25.21254342v1

Epitweetr evaluation: Manual vs epitweetr

Table 2. Summarised metrics of epitweetr evaluation.

Variables	Manual method	epitweetr
Number of signals	188	448
Number of events	184	334
IRA [95% CI]	47.9% [43.8% - 52.0%]	78.6% [75.2% - 82.0%]
PPV [95% CI]	97.9% [95.8% - 99.9%]	74.6% [70.5% - 78.6%]

Figure 4. Histogram showing validation time differences in hours between epitweetr and manual method

Source: https://www.medrxiv.org/content/10.1101/2021.03.25.21254342v1