A Customisable Pipeline for Continuously Harvesting Socially-Minded Twitter Users

Paolo Missier(*), Alexander Romanovsky(*), Nélio Cacho(+), Flavio Primo(*), Mickael Figueredo(+)

(*) Newcastle University, UK
(+) Federal University of Rio Grande do Norte, Brazil
Motivation

Problem: Zika, Dengue, Chikungunya are dangerous endemic diseases in Brazil

Solution?

- Government has low resources
- Crowdsourcing has "needle in a haystack" problem
Goal: identify target users for recruitment

**Activist**: a person who demonstrates an inclination to become engaged in social issues, regardless of the specific topic

- “socially-minded”

Can it be…

- Defined with computable metrics?
- Discovered independently from its social focus?

Customisable pipeline for continuous discovery of users of interest in multiple topics in social media
Approach

**Goal**: find users actively engaged in events

Defined by → **user metrics**
Measurements of relationships and social activities over Twitter

Defined by → **contexts**
Queries that retrieve event-related content
**Contexts**

\[ C_{\text{context}} = (K_{\text{hashtags}}, \Delta t_{\text{date interval}}, s_{\text{location}}) \]

\[ P(C) \quad \text{set of posts} \rightarrow \quad p \in P(C) \quad \text{post} \quad u(p) \in P(C) \quad \text{author of post} \]

\[ \tilde{P}(C) = P(C') \setminus P(C) \quad \text{complement of } P(C) \]

\[ G_C = (V, E) \rightarrow \quad V = \{ u(p) | p \in P(C') \} \quad \text{set of nodes} \]

\[ e = \langle u_1_{\text{post author}}, u_2_{\text{mentioned/retweeted author}}, w_{\text{weight}} \rangle \quad \text{edge} \]
1. Harvesting content from context

Contexts

C_1

...  

Twitter API

S_1

...  

Twitter streams

Context

Date range  

Hashtags  

Location

Stream

tweet_1 { 
author mentions
...
}
2. Network creation

Twitter Streams

Create network

Network Graphs

Relations network
- directed
- weighted
3. Community detection

Network Graphs

\[ G_1, \ldots, G_n \]

Partitioned Network Graphs

\[ PG_1, \ldots, PG_n \]

Infomap

Communities of users

- inner high information flow
- remove communities <3

Network Graphs connected to Infomap, resulting in Partitioned Network Graphs.
3. Community detection: Infomap

Random Walker path:

Huffman Coding of the nodes

Random Walker path:

Code length: 314 bits
3. Community detection: Infomap

Random Walker path:

111 0000 11 01 101 100 101 01 0
001 0 110 011 00 110 00 111 101
1 10 111 00 01 011 00 111 01 0
011 10 000 111 10 111 10 0010 1
0 011 010 011 10 000 111 0001 0
111 010 100 011 00 111 00 011 0
1 011 00 111 110 111 110 1011 1
11 01 101 01 0001 0 110 111 00
011 110 111 1011 10 111 000 10
000 111 0001 0 111 010 1010 01
0 1011 110 00 10 011

Code length: 243 bits

Dual problem:
detect communities by compressing the description of information flows on networks
4. Profiling

Characterize users with metrics

**Context independent**
- Follower rank

**Context specific**
- Indegree centrality

**Content based**
- Topical Focus
- Topical Strength
- Topical Attachment
4. Profiling

Context independent:

\[
FR(u) = \frac{|\text{followers}(u)|}{|\text{followers}(u)| + |\text{following}(u)|}
\]

Context specific:

\[
IC(u) = \frac{\text{indegree}(u)}{N - 1}
\]
4. Profiling

Content based:

\[
\begin{align*}
TF(u)_{TopicalFocus} &= \frac{P1_{on}(u)}{P1_{off}(u) + 1} \\
TA(u)_{TopicalAttachment} &= \frac{P1_{on}(u) + P2_{on}(u)}{P1_{off}(u) + P2_{off}(u) + 1} \\
TS(u)_{TopicalStrength} &= \frac{P2_{on}(u) \cdot \log(P2_{on}(u) + R3_{on} + 1)}{P2_{off}(u) \cdot \log(P2_{off}(u) + R3_{off} + 1) + 1}
\end{align*}
\]

Where:

- P1: # of original posts by u in C
- P2: # urls found in original posts by u in C
- R3: # of retweets of u's tweets
5. Ranking

\[ R1(u) = \frac{1}{\sum_{u \in C} IC(u) + 1} \cdot \sum_{u \in C} TF(u) \]

↑ on topic ↓ community leader

\[ R2(u) = |FR(u) - 1| \cdot \left( \sum_{u \in C} TA(U) + \sum_{u \in C} IC(U) \right) \]

↑ on topic ↑ community leader ↓ popularity

\[ R3(u) = |FR(u) - 1| \cdot \left( \sum_{u \in C} TA(U) + \frac{1}{\sum_{u \in C} IC(U) + 1} \right) \]

↑ on topic ↓ community leader ↓ popularity
The complete user extraction pipeline

1. Bootstrap context list
2. Context detection
3. Network creation
4. Community detection
5. Profiling
6. Ranking
7. New context detection

Twitter
Query
Context metadata
Tw eets
Network graph
User communities
Profiles
Context metadata
User timelines
Profiles DB
Features
Ranked user list
Evaluation: tweets harvesting

Unsupervised validation method
A posteriori validation of the generated user ranking

no prior ground truth
**Evaluation: test contexts**

| Context name                                      | Period (2018)   | Nodes | Edges | Density | Avg degree | Assortativity |
|---------------------------------------------------|-----------------|-------|-------|---------|------------|---------------|
| 16 days of action                                 | 11-25 / 12-10   | 396   | 349   | 0.002   | 1.8        | -0.1          |
| Elf day                                           | 12-03 / 12-12   | 365   | 436   | 0.003   | 2.4        | -0.2          |
| Dry January                                       | 01-01 / 01-31   | 235   | 234   | 0.004   | 2.0        | -0.3          |
| Cervical cancer prevention week                   | 01-21 / 01-27   | 209   | 192   | 0.004   | 1.8        | -0.1          |
| Time to talk day                                  | 02-06 / 02-07   | 268   | 231   | 0.003   | 1.7        | -0.2          |
| Eating disorder awareness week                    | 02-25 / 03-02   | 256   | 241   | 0.004   | 1.9        | -0.2          |
| Rare disease day                                  | 02-28 / 03-01   | 294   | 206   | 0.002   | 1.4        | -0.2          |
| Ovarian cancer awareness month                    | 03-01 / 03-31   | 215   | 202   | 0.004   | 1.9        | -0.4          |
| Nutrition and hydration week                      | 03-11 / 03-17   | 273   | 326   | 0.004   | 2.4        | -0.3          |
| Brain awareness week                              | 03-11 / 03-17   | 307   | 281   | 0.003   | 1.8        | -0.1          |
| No smoking day                                    | 03-13 / 03-14   | 254   | 219   | 0.003   | 1.7        | -0.3          |
| Epilepsy awareness purple day                     | 03-26 / 03-27   | 306   | 252   | 0.003   | 1.6        | -0.2          |
| Experience of care week                           | 04-23 / 04-27   | 176   | 196   | 0.006   | 2.2        | -0.1          |
| Brain injury week                                 | 05-01 / 05-31   | 238   | 306   | 0.005   | 2.6        | -0.1          |
| Mental health awareness week                      | 05-14 / 05-20   | 268   | 245   | 0.003   | 1.8        | -0.5          |
| Dementia action week                              | 05-21 / 05-31   | 300   | 300   | 0.003   | 2.0        | -0.0          |
| Mnd awareness month                               | 06-01 / 06-30   | 141   | 234   | 0.012   | 3.3        | -0.3          |
| Wear purple for jia                               | 06-01 / 06-30   | 165   | 245   | 0.009   | 3.0        | -0.5          |
| Carers week                                       | 06-11 / 06-17   | 270   | 277   | 0.004   | 2.1        | 0.0           |
| National dementia carers                          | 09-09 / 09-10   | 184   | 177   | 0.005   | 1.9        | -0.2          |
| Mens health week                                  | 06-11 / 06-17   | 264   | 214   | 0.003   | 1.6        | -0.2          |
| Stress awareness day                              | 11-07 / 11-08   | 293   | 209   | 0.002   | 1.4        | -0.2          |
| National dyslexia week                            | 10-01 / 10-07   | 229   | 235   | 0.004   | 2.1        | -0.2          |
| Ocd awareness week                                | 10-07 / 10-13   | 202   | 193   | 0.005   | 1.9        | -0.6          |
| Jeans for genes day                               | 09-21 / 09-22   | 246   | 325   | 0.005   | 2.6        | -0.2          |

- **25 contexts**
- **1 day → 1 month**
- **254 nodes**
- **235 edges**
- **2 density**
- **-0.2 assortativity**
Evaluation: repeated users

Repeated users are ranked higher (multiple participations to contexts).

example:

\[ \sum_{u \in C} TF(u) \]

Remove inactive users:

- \( FR(u) = 0 \)
- \( \text{min}_\text{max}(|\text{Tweets}(u)|) < 0.005 \)
## Evaluation

| Username       | Name                                | Follower rank | Participations |
|----------------|-------------------------------------|---------------|----------------|
| alzheimerssoc  | Alzheimer’s Society                 | 0.99          | 4              |
| dementiauk     | Dementia UK                         | 0.98          | 4              |
| mentalhealth   | Mental Health Fdn                   | 0.97          | 3              |
| colesmillerllp | Coles Miller LLP                    | 0.65          | 3              |
| jeremy_hunt    | Jeremy Hunt                         | 1.0           | 2              |
| nhsengland     | NHS England                         | 0.99          | 2              |
| carersuk       | Carers UK                           | 0.95          | 2              |
| rdash_nhs      | RDaSH NHS FT                        | 0.88          | 2              |
| alzsocseeengland | Alzheimer’s Society - South ... | 0.64          | 2              |
| mndassoc       | MND Association                     | 0.64          | 2              |

**Top 10 repeat users**
Evaluation

Number of repeat users for each context
Evaluation: choice of ranking function

Number of ranked users: 3567

R3 effectively finds individuals
## Evaluation

|   #  | User            | Ranking 1     |          | User            | Ranking 2     |          | User            | Ranking 3     |          |
|------|-----------------|---------------|----------|-----------------|---------------|----------|-----------------|---------------|----------|
|      |                 | On-topic      | Individual |                 | On-topic      | Individual |                 | On-topic      | Individual |
| 1    | homesnutrition  | X             |          | johnneustadt    | X             |          | johnneustadt    | X             |          |
| 2    | ficajones       | X             | X        | jo_millar27     | X             | X        | solutions777    | X             | X        |
| 3    | helenvweaver    | X             | X        | hatchbrenner    |               |          | kingste29344921 | X             | X        |
| 4    | spriggsnutri    | X             |          | nchawkes        | X             | X        | daisylu1964     |               |          |
| 5    | critcarelthtr   | X             |          | moz0373runner   | X             | X        | zakariamarsli   | X             | X        |
| 6    | danielleroisin_ | X             | X        | aimsonhealth    | X             | X        | meowaaaaaa      |               |          |
| 7    | mynameisandyj   | X             | X        | wordsharkv5     |               | X        | vecta67         |               | X        |
| 8    | fionaliu92      | X             | X        | fullcircle_play |               |          | cosfordfamily1  | X             | X        |
| 9    | ldppartnership  | X             |          | qsprivatehealth | X             |          | hayleycorriganx |               | X        |
| 10   | milaestevam1    | X             |          | socialissp      |               |          | jhbrasie        |               | X        |

Top 10 ranked users for the ranking functions R1, R2 and R3
What are we doing now

Recursive Twitter contexts expansion

Users-hashtags clustering

Context detection automation
Thank you!

A Customisable Pipeline for Continuously Harvesting Socially Minded Twitter Users

Paolo Missier(*), Alexander Romanovsky(*), Nélio Cacho(+), Flavio Primo(*), Mickael Figueredo(+)

Slides, code and full-text paper available @ https://flavioprimo.xyz/blog/a-customisable-pipeline-for-continuously-harvesting-socially-minded-twitter-user/

Email: fla.primo.engineer@outlook.com Twitter: @flavioprimo_91