SciencesPo

Mine it, mine all!

Digital methods for mining and measuring the spread of fact-checked claims on social media, featuring the open-source tools `minet` and `minall`

Kelly Christensen

médialab
Sciences Po
26 February 2024

DE FACTO
Des clés pour mieux s'informer

SciencesPo
MÉDIALAB

De Facto Media Observatory

De Facto is France's EDMO hub, supported (2021-2023) by the following partners:

▶ AFP : fact-checkers

▶ médialab : researchers (my group!)

▶ CLEMI : media literacy

▶ XWiki : database / website

Fact-checkers contribute new reviews to the De Facto database almost every day. Every night, those reviews are enriched with up-to-date metrics.

The database's enrichment process includes the following:

▶ Research where the fact check was shared and how many people engaged with it online.

▶ Extract the content and metadata about appearances of the fact-checked claim.

Terminology Note: "fact check" > "fake news"

Not everything that is reviewed is determined to be fake. Sometimes fact checks confirm the truth of a claim. A "fact check" is not the same thing as "fake news."

Outline

Fact Checks : Key Terms

Fact-checkers **review** a **claim**, which made an **appearance** online.

The Review

> An article describing and evaluating the accuracy of a claim that someone made publicly.
Example: Politician Z claims X caused Y, which is not true.

The Claim

> A thesis asserting something about an event, a person, etc.
Example: X caused Y.

The Appearance

> The physical or digital material through which the claim was expressed and/or received by people.
Example: The video on YouTube of politician Z's public speech.

ClaimReview

In its database, the De Facto hub organizes this information using the international standard *ClaimReview*.

ClaimReview is part of *Schema.Org*, a set of vocabularies used to structure nested data. Founded by Google, Microsoft, Yahoo and Yandex, the *Schema.Org* vocabularies are managed by a community of developers and organize data all over the internet.

In example, let's organize data about the conference location:
- ▶ @type : *CollegeOrUniversity*
- ▶ Name: *University of Amsterdam*
- ▶ Logo:
    - ▶ @type : *ImageObject*
    - ▶ Width : 500
    - ▶ ContentUrl: https://www.uva.nl/binaries/...

UNIVERSITEIT
VAN AMSTERDAM

The *ClaimReview* format organizes information about a fact check. Descending from the root, there are a few key branches.

- ▶ itemReviewed (`@type = "Claim"`)
    - ▶ **appearance** (e.g. `@type = "SocialMediaPosting"`)
        - ▶ url (e.g. `"url"`: `"https://twitter.com/USER/status/POST"`)
- ▶ author (e.g. `@type = "Organization"`)
- ▶ **reviewRating** (e.g. `@type = "Rating"`)
    - ▶ ratingValue (e.g. `2`)
    - ▶ alternateName (e.g. `Missing Context`)
- ▶ **claimReviewed** (e.g. *A photo of person A after event B.*)

ClaimReview

The **appearance** information can be very detailed.

- ▶ "@type": "SocialMediaPosting"
- ▶ "url": "https://twitter.com/USER/status/POST"
- ▶ "interactionStatistic":
  - ▶ "interactionType": "Share Action"
  - ▶ "interactionService"
    - ▶ @type: "Website"
    - ▶ "name": "Twitter"
  - ▶ "userInteractionCount": 2513
- ▶ "sharedContent":
  - ▶ "@type": "ImageObject"
  - ▶ "contentURL":
    "https://pbs.twimg.com/media/IDENTIFIER.jpg"
- ▶ "text": *[Text in the tweet, which accompanies the shared image.]*

Occasionally, we've needed to further develop the *ClaimReview* standard to better describe online appearances of claims.

For example, sometimes the types of ways users interact with the claim's online appearance are not among the types available in *ClaimReview*. We've created a new type, "Engagement."

```
"interactionStatistic":
    ▶ "interactionType":  "defacto:EngagementAction"
    ▶ "interactionService"
        ▶ @type:  "Website"
        ▶ "name":  "Reddit"
    ▶ "userInteractionCount":  17
```

Improving ClaimReview: New Creator Type

In *ClaimReview*, the author of an **appearance** ("*Creator*") can either be a person or an organization.

However, sometimes a social media account, especially a bot or an anonymous account, is ontologically neither a person nor an organization. Furthermore, accounts can have distinctive properties.

```
"creator":
```
▶ `"@type":` `"defacto:SocialMediaAccount"`
▶ `"defacto:dateCreated":` `"2022-03-07T10:23:41"`
▶ `"identifier":` `"12345"`
▶ `"name":` *Username*
▶ `"interactionStatistic"`
   ▶ `"interactionType":` `"FollowAction"`
   ▶ `"interactionService"`
      ▶ `@type:` `"Website"`
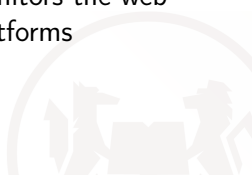   ▶ `"name":` `"Twitter"`
   ▶ `"userInteractionCount":` `21344`

A *ClaimReview* includes one or more **appearances** of a **reviewed claim**.

The **appearance**, by virtue of being something seen online, includes a URL.

And we can do a lot with a URL!

▶ See where it came from (domain)

▶ Scrape its content

▶ Search for it in a database (Buzzsumo) that monitors the web and tracks when a URL appears on different platforms

With just a URL (and API keys), the open-source tool `minet`[1] can extract data from webpages, Facebook, Meta's CrowdTangle API, YouTube, Twitter, etc.

Knowing what type of URL you have (i.e. a YouTube link), you can call one of `minet`'s dedicated commands (i.e. `minet youtube`).

Challenge

When fact-checkers share *ClaimReviews* with the De Facto hub, they're combining URLs from all over the internet, requiring an undetermined variety of `minet` commands.

[1]https://github.com/medialab/minet

`minall`[2] - Mine all of it!

Challenge 1

Variety of URLs, requiring different data mining methods.

Solution

Interpret the needs of each URL and process it with the necessary `minet` commands, but in a batched and multi-threaded way.

Challenge 2

Things on the internet disappear. Update without data loss.

Solution

Accept a set of URLs **and** a previous iteration of enrichment. Using an in-memory SQL database, only update/overwrite the URL's metadata if the new data extraction was successful.

[2]https://github.com/medialab/minall

`minall` - Mine all of it!

See the tutorial online: https://medialab.github.io/minall/tutorials/

What we start with:

▶ API keys (get a YouTube API key for free from Google):

```
---
youtube:
    key: "XXXX"
```

▶ URLs from fact-checked claims

rating	url
True	www.online-news.eu/story
Missing Context	www.youtube.com/watch/video

minall - Mine all of it!

▶ `config.yml` (API Configuration)

```
---
youtube:
    key: "XXXX"
```

▶ `data.csv` (URLs from fact-checked claims)

rating	url
True	www.online-news.eu/story
Missing Context	www.youtube.com/watch/video

```
minall --config config.yml --output-dir output/
        --links data.csv --url-col url
```

minall - Mine all of it!

SciencesPo
MÉDIALAB

- ▶ Data fields we start with
  - ▶ url
- ▶ Data fields we end up with (selected columns)

  - ▶ url
  - ▶ date_published
  - ▶ domain
  - ▶ work_type
  - ▶ duration
  - ▶ identifier
  - ▶ date_modified
  - ▶ country_of_origin
  - ▶ abstract
  - ▶ keywords
  - ▶ title
  - ▶ text
  - ▶ hashtags
  - ▶ facebook_comment

  - ▶ facebook_like
  - ▶ facebook_share
  - ▶ pinterest_share
  - ▶ twitter_share
  - ▶ twitter_like
  - ▶ tiktok_share
  - ▶ tiktok_comment
  - ▶ reddit_engagement
  - ▶ youtube_watch
  - ▶ youtube_comment
  - ▶ youtube_like
 - ▶ youtube_favorite
  - ▶ youtube_subscribe

title

- Data fields we start with
  - url
- Data fields we end up with (selected columns)

  - ...
  - creator_type
  - creator_date_created
  - creator_location_created
  - creator_identifier
  - creator_facebook_subscribe

  - creator_facebook_follow
  - creator_twitter_follow
  - creator_youtube_subscribe
  - creator_create_video
  - creator_name
  - creator_url
  - create_video

By extracting counts of likes and followers, we can explore possible
correlations between the size of a user's account and the number of
likes their posts get on different platforms in our database.
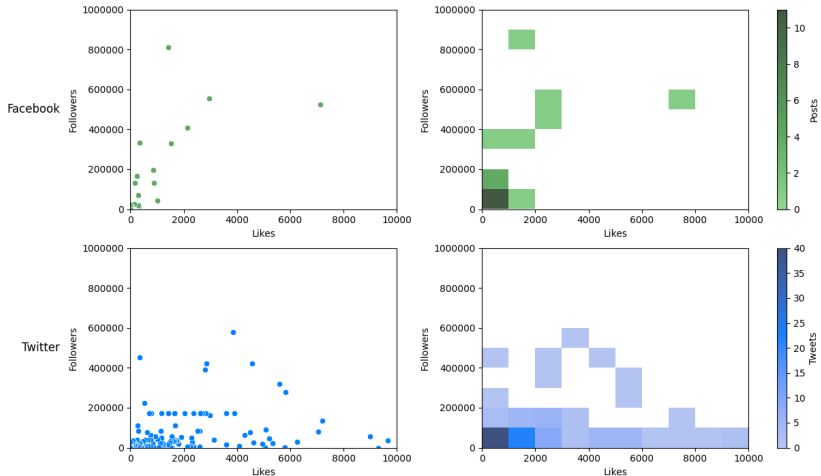
We can also combine discrete and categorical data. For example, we can examine how many of an account's followers liked a post when it was rated false (1, magenta) or partly false (2, violet).

Facebook · Twitter
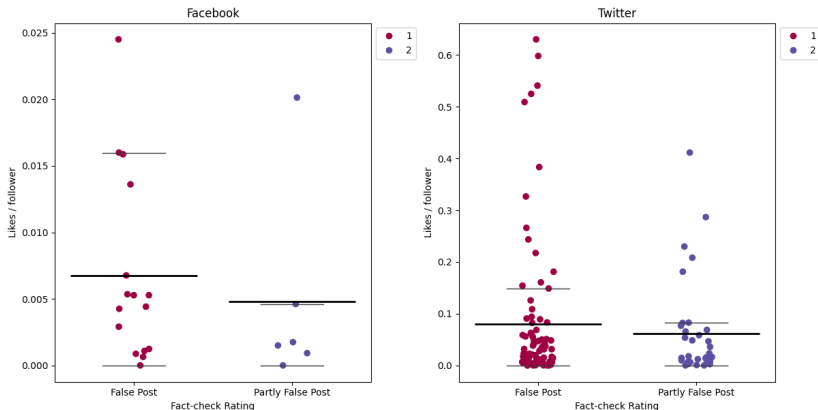
3For Twitter dataset, when likes / follower > 0.9, the outlier was removed. Removed 4 outliers.

What we can do with the new info

Finally, we can explore the text data that we extracted from various parts of the fact-checked URLs' web content.

Tweet: Post

Coup d'Etat: Macky Sall, marionnette de Macron , annule les élections au Sénégal...

Webpage: Article Text

... Darmanin, le « chouchou » de Macron à quitte ou double...

YouTube Video: Keywords

Emmanuel Macron, Président République Française , France, Elysée, Présidence

Conclusions

1. We start with what the fact-checkers provide, URLs.
2. We jump from those URLs into the internet, to extract the content and discover the links' circulation and engagement online.
   ▶ Through an automated and free (GitHub CI) processs, we update our enriched *ClaimReviews* every day.
3. We aggregate that new information and try to better understand the nature of fact-checked content online.

All of this enrichment is free to do (minus the cost of API keys). The tools are transparent and published in public repositories on GitHub.[4]

---

[4]https://github.com/medialab/minet, https://github.com/medialab/minall, https://github.com/medialab/defacto-enrichment-public