Analysis of Facebook (Public) Conversations: Prospects and Challenges

Matteo Tarantino

AI & Policy Talks – Dec 3, 2020

Research Context

- Part of larger COVI_MEDIA project with 8 research units per country (ITA & US) studying **media-related phenomena about the first 90 days** of the pandemic (Jan 1° - April 30° 2020).
- My unit: focus on **bottom-up discourse on social media** about COVID-19.
    - What are the main discursive modes about COVID-19?
    - What kind of users are active in COVID-19 topics at different stages?

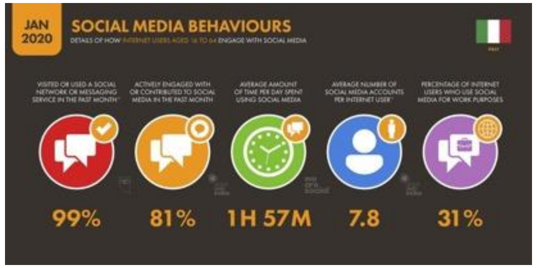
Objective

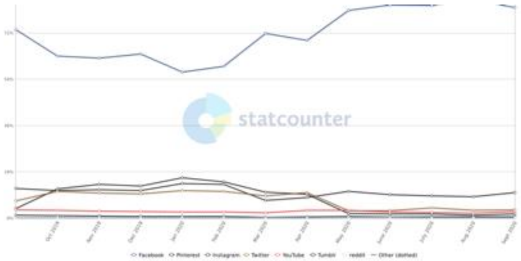
- Understand the topics, tones & practices shaping public discourse on social media in Italy during the initial phase of the COVID-19 pandemic.
- Underlying assumption: social media as proxy for social discourse.
- Research setting: **Facebook**
  - **Twitter**: easier to do data collection, but skewed socio-demographically; **Instagram**: more difficult to do data collection, less and skewed socio-demographically.

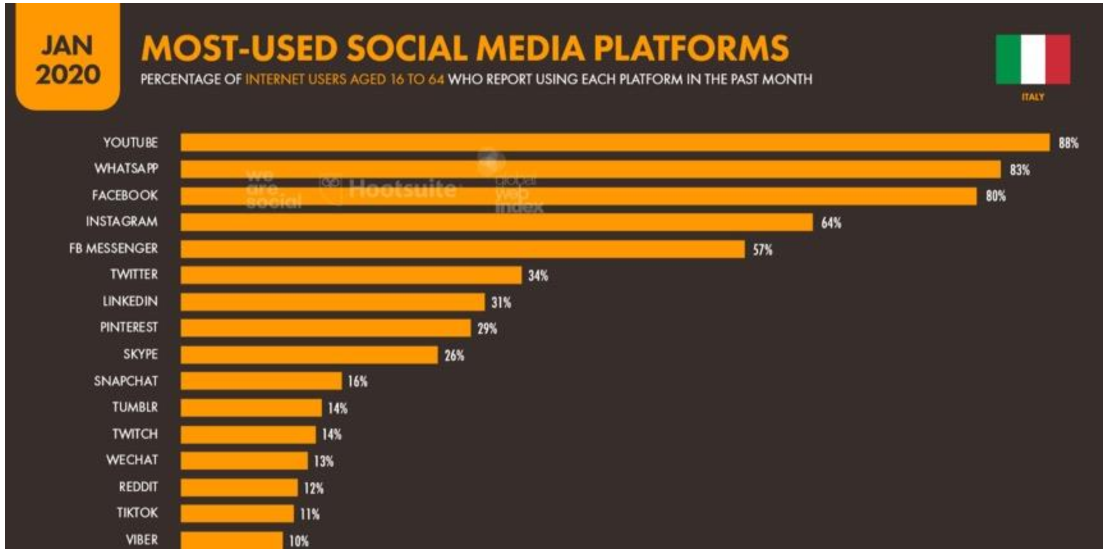
Methods

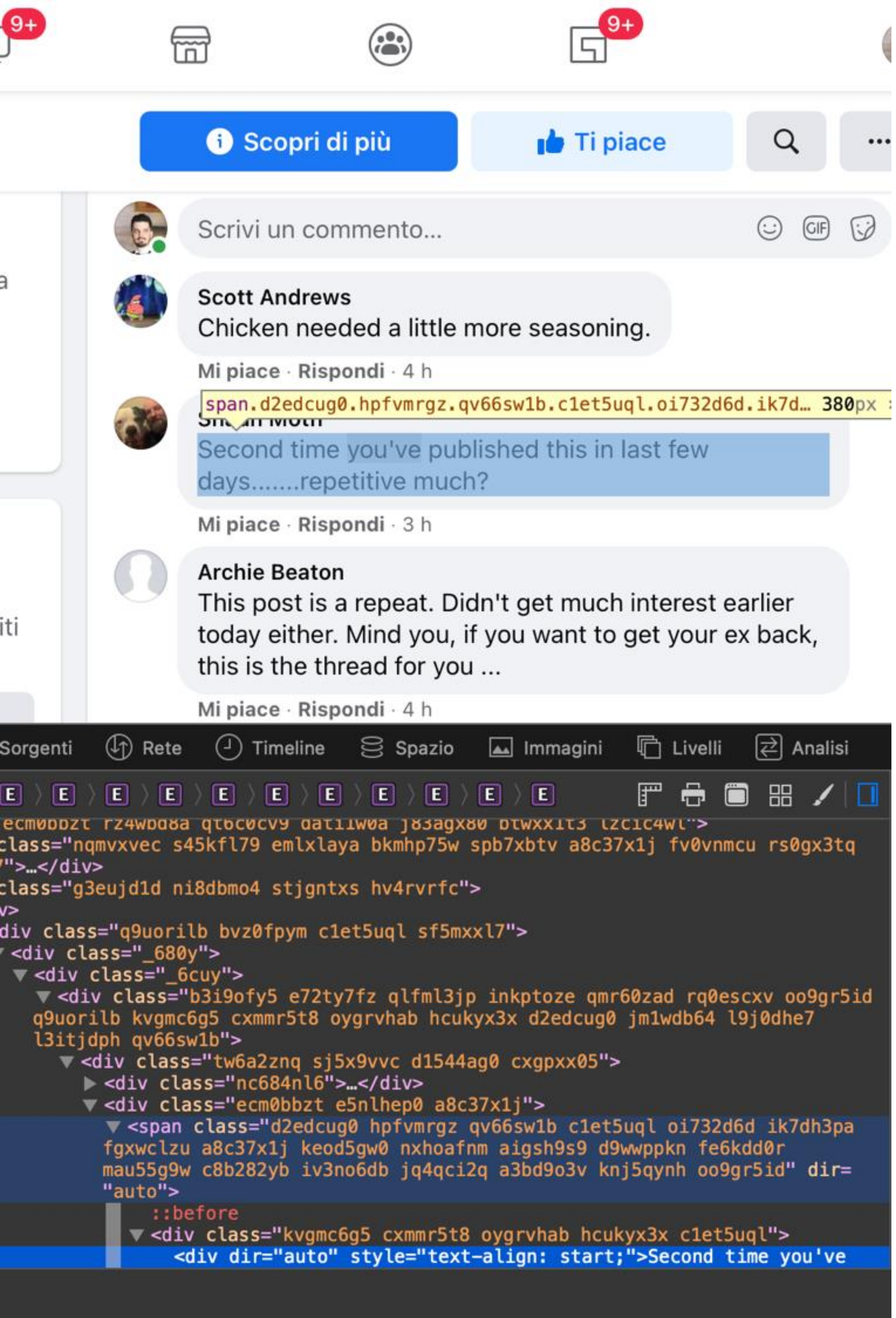
- **Corpus**: all text-containing comments on all posts published on COVID between January 1° and March 31° (n=2,368) by the most popular newspaper FB account, *La Repubblica* (n=705,538).

- **Main Techniques**:
  - Analysis of content (word frequencies, bigrams & trigrams, collocations + manual coding) → identification of topics
  - Analysis of users behaviors (frequency, quantity and length of posting, lexical variation, content) → User profiles and patterns









Scott Andrews
Chicken needed a little more seasoning.
Mi piace · Rispondi · 4 h

span.d2edcug0.hpfvmrgz.qv66sw1b.c1et5uql.oi732d6d.ik7d… 380px

Second time you've published this in last few days.......repetitive much?
Mi piace · Rispondi · 3 h

Archie Beaton
This post is a repeat. Didn't get much interest earlier today either. Mind you, if you want to get your ex back, this is the thread for you ...
Mi piace · Rispondi · 4 h

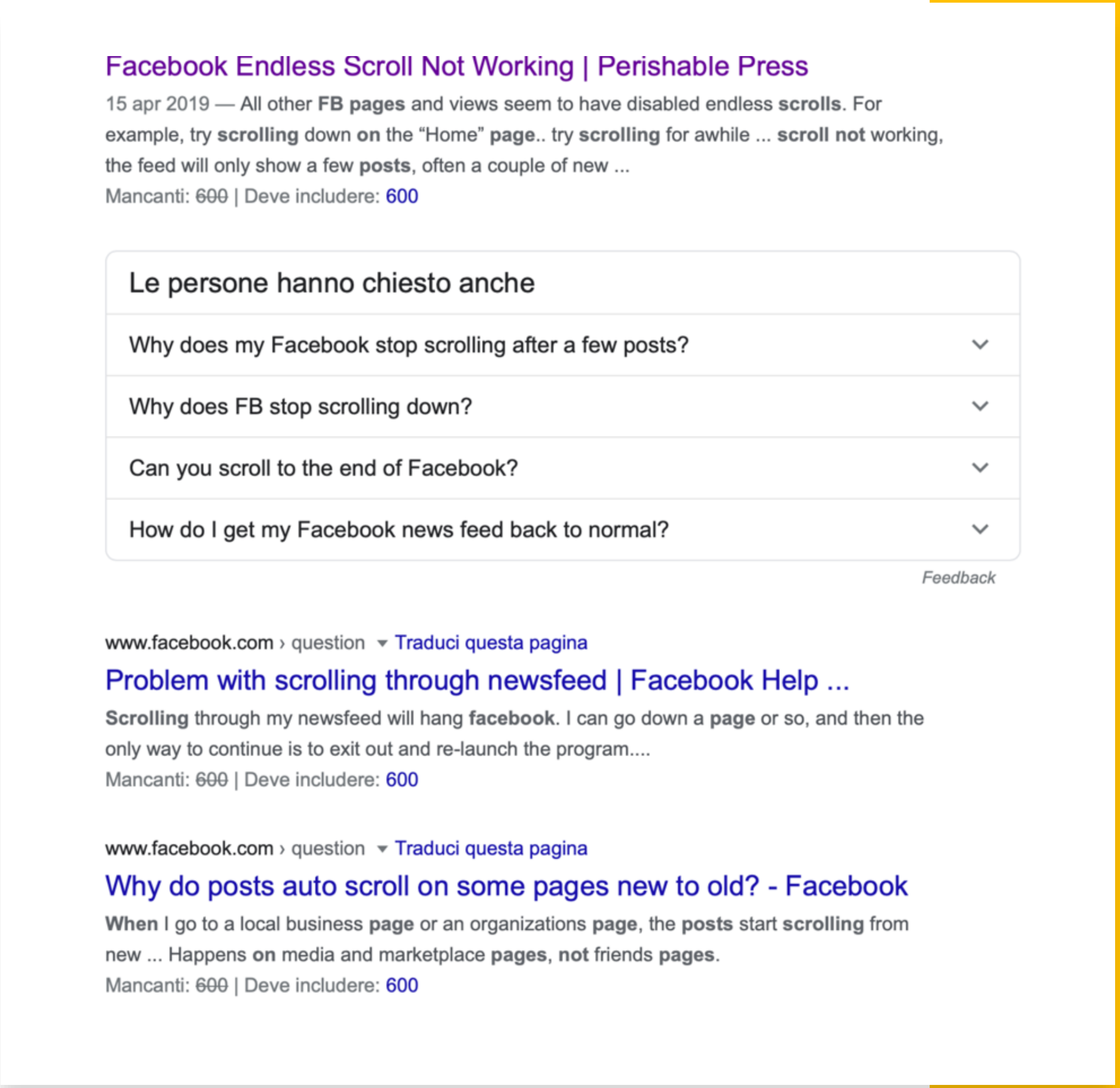
```
ecm0bbzt rz4wbd8a qt6c0cv9 datilw0a j83agx80 btwxx1t3 lzcic4wl">
class="nqmvxvec s45kfl79 emlxlaya bkmhp75w spb7xbtv a8c37x1j fv0vnmcu rs0gx3tq
">…</div>
class="g3eujd1d ni8dbmo4 stjgntxs hv4rvrfc">
v>
div class="q9uorilb bvz0fpym c1et5uql sf5mxxl7">
 <div class="_680y">
  ▼ <div class="_6cuy">
    ▼ <div class="b3i9ofy5 e72ty7fz qlfml3jp inkptoze qmr60zad rq0escxv oo9gr5id
     q9uorilb kvgmc6g5 cxmmr5t8 oygrvhab hcukyx3x d2edcug0 jm1wdb64 l9j0dhe7
     l3itjdph qv66sw1b">
      ▼ <div class="tw6a2znq sj5x9vvc d1544ag0 cxgpxx05">
        ▶ <div class="nc684nl6">…</div>
        ▼ <div class="ecm0bbzt e5nlhep0 a8c37x1j">
          ▼ <span class="d2edcug0 hpfvmrgz qv66sw1b c1et5uql oi732d6d ik7dh3pa
            fgxwclzu a8c37x1j keod5gw0 nxhoafnm aigsh9s9 d9wwppkn fe6kdd0r
            mau55g9w c8b282yb iv3no6db jq4qci2q a3bd9o3v knj5qynh oo9gr5id" dir=
            "auto">
              ::before
            ▼ <div class="kvgmc6g5 cxmmr5t8 oygrvhab hcukyx3x c1et5uql">
                <div dir="auto" style="text-align: start;">Second time you've
```

The infrastructure

- Post-Cambridge Analytica, Facebook Graph API is (understandably) increasingly limited for these kinds of analyses: **a post-API World** (Freelon, 2018).
- Collecting Facebook conversation requires the development (and maintenance) of a dedicated infrastructure which mimicks an user and is able to:
  - Collect the posts published by a source
  - Collect each comment published underneath a post
  - Collect replies to that comment
- The core technique is html parsing:
  - Content is loaded and pertinent html elements (e.g. the text of a comment) are identified by their css identifiers.

The problem of sourcing

- Getting the posts published by a public page is limited by Facebook stopping scrolling after a few thousand posts.
- This forces the infrastructure to involve a third-party provider who archives Facebook posts from pages.
    - Problem of interoperability



Facebook Endless Scroll Not Working | Perishable Press

15 apr 2019 — All other **FB pages** and views seem to have disabled endless **scrolls**. For example, try **scrolling** down **on** the "Home" **page**.. try **scrolling** for awhile ... **scroll not** working, the feed will only show a few **posts**, often a couple of new ...

Mancanti: ~~600~~ | Deve includere: 600

Le persone hanno chiesto anche

Why does my Facebook stop scrolling after a few posts?

Why does FB stop scrolling down?

Can you scroll to the end of Facebook?

How do I get my Facebook news feed back to normal?

Feedback

www.facebook.com › question ▾ Traduci questa pagina

Problem with scrolling through newsfeed | Facebook Help ...

Scrolling through my newsfeed will hang **facebook**. I can go down a **page** or so, and then the only way to continue is to exit out and re-launch the program....

Mancanti: ~~600~~ | Deve includere: 600

www.facebook.com › question ▾ Traduci questa pagina

Why do posts auto scroll on some pages new to old? - Facebook

When I go to a local business **page** or an organizations **page**, the **posts** start **scrolling** from new ... Happens **on** media and marketplace **pages**, **not** friends **pages**.

Mancanti: ~~600~~ | Deve includere: 600

Blindfolded Chess

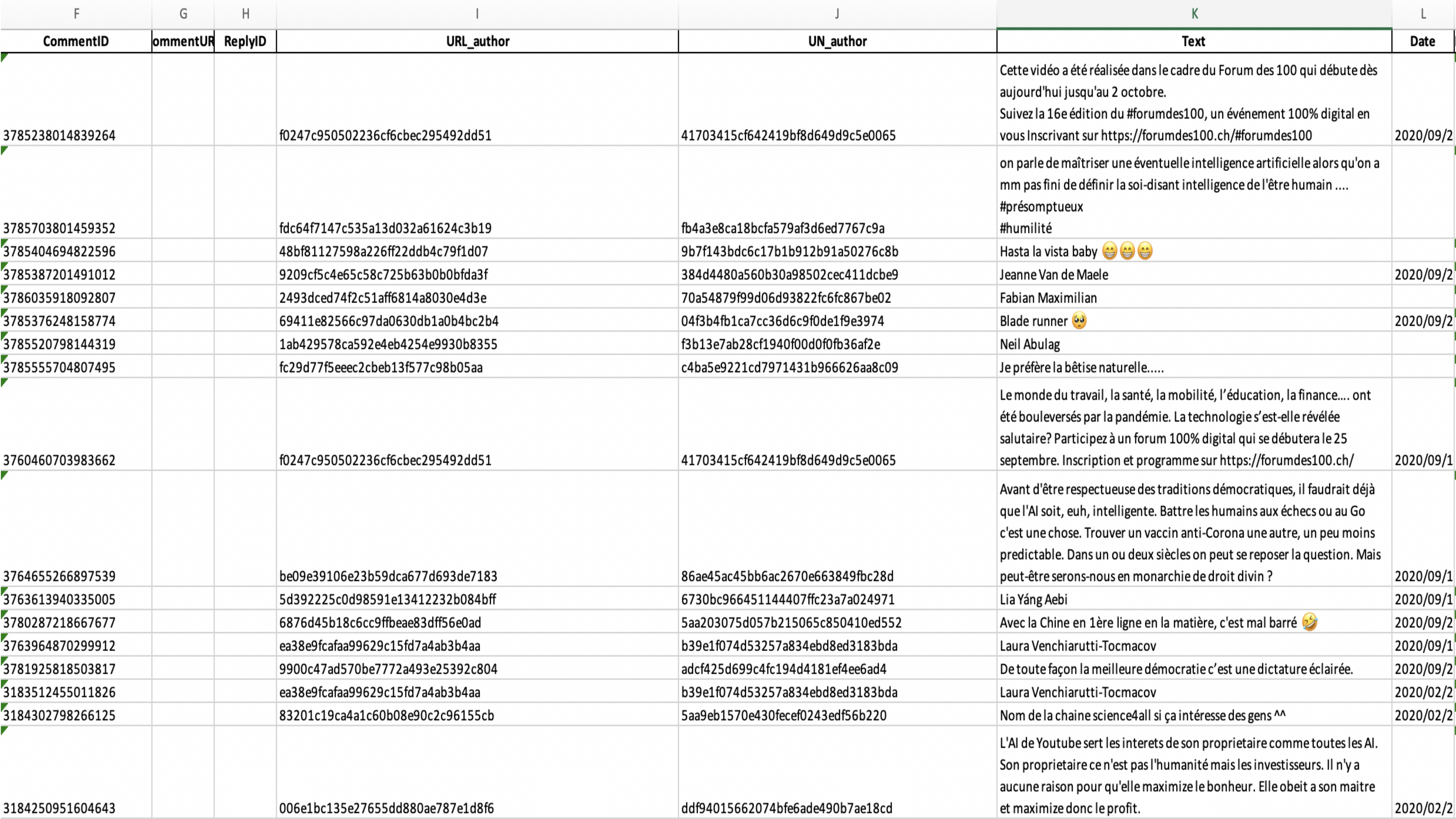
- Facebook protects itself using:
  - TOS specifications.
  - “Smart” detection techniques against harvesting agents, based for example on the frequency of requests, user-agents etc. leading to bans.
  - “Dumb” Interface limitations (i.e. scrolls are limited)
  - “Mid-level” limitations such as frequent css reshuffling which confuses the harvesting agent.
- The agent does not know when and what changes will be operated by Facebook
  - Constant, expensive maintenance.

Privacy Protection

- Identification of users for statistical purposes requires unique identifiers (user profile url)

- Ethical and legal challenge of privacy and compliance (GDPR compliance).
  - Academic research usually falls under «public interest» for GDPR purposes.

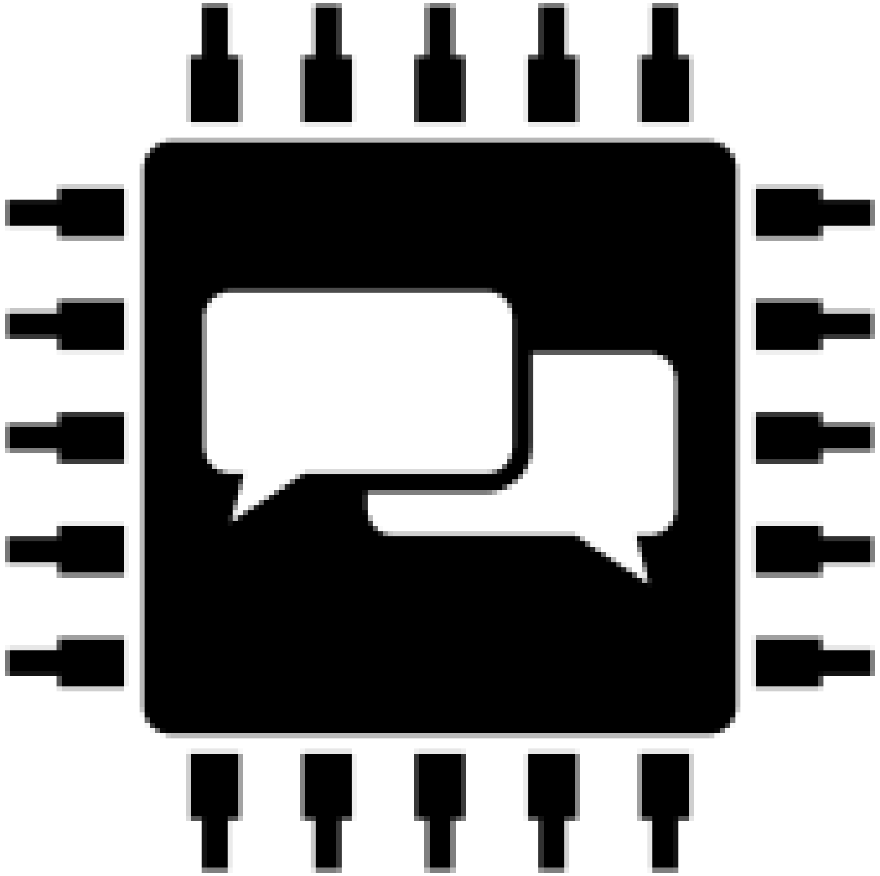
- We adopted anonymization through SHA256 encryption **<u>at the source</u>** of usernames and profile urls, which we used as user indentifiers.



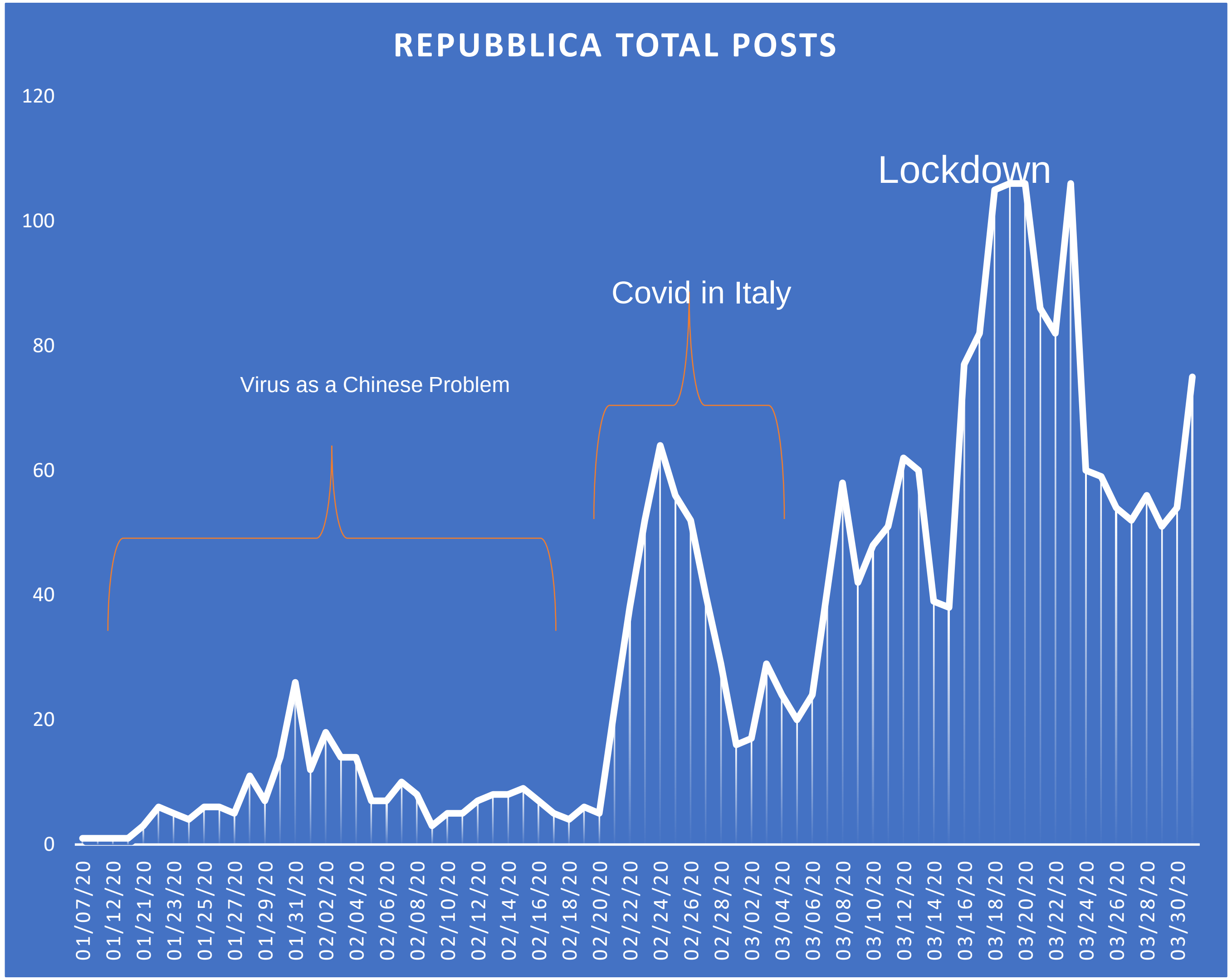
F	G	H	I	J	K	L
CommentID	ommentUR	ReplyID	URL_author	UN_author	Text	Date
3785238014839264			f0247c950502236cf6cbec295492dd51	41703415cf642419bf8d649d9c5e0065	Cette vidéo a été réalisée dans le cadre du Forum des 100 qui débute dès aujourd'hui jusqu'au 2 octobre. Suivez la 16e édition du #forumdes100, un événement 100% digital en vous Inscrivant sur https://forumdes100.ch/#forumdes100	2020/09/2
3785703801459352			fdc64f7147c535a13d032a61624c3b19	fb4a3e8ca18bcfa579af3d6ed7767c9a	on parle de maîtriser une éventuelle intelligence artificielle alors qu'on a mm pas fini de définir la soi-disant intelligence de l'être humain .... #présomptueux #humilité	
3785404694822596			48bf81127598a226ff22ddb4c79f1d07	9b7f143bdc6c17b1b912b91a50276c8b	Hasta la vista baby 😁😁😁	
3785387201491012			9209cf5c4e65c58c725b63b0b0bfda3f	384d4480a560b30a98502cec411dcbe9	Jeanne Van de Maele	2020/09/2
3786035918092807			2493dced74f2c51aff6814a8030e4d3e	70a54879f99d06d93822fc6fc867be02	Fabian Maximilian	
3785376248158774			69411e82566c97da0630db1a0b4bc2b4	04f3b4fb1ca7cc36d6c9f0de1f9e3974	Blade runner 🥺	2020/09/2
3785520798144319			1ab429578ca592e4eb4254e9930b8355	f3b13e7ab28cf1940f00d0f0fb36af2e	Neil Abulag	
3785555704807495			fc29d77f5eeec2cbeb13f577c98b05aa	c4ba5e9221cd7971431b966626aa8c09	Je préfère la bêtise naturelle.....	
3760460703983662			f0247c950502236cf6cbec295492dd51	41703415cf642419bf8d649d9c5e0065	Le monde du travail, la santé, la mobilité, l'éducation, la finance.... ont été bouleversés par la pandémie. La technologie s'est-elle révélée salutaire? Participez à un forum 100% digital qui se débutera le 25 septembre. Inscription et programme sur https://forumdes100.ch/	2020/09/1
3764655266897539			be09e39106e23b59dca677d693de7183	86ae45ac45bb6ac2670e663849fbc28d	Avant d'être respectueuse des traditions démocratiques, il faudrait déjà que l'AI soit, euh, intelligente. Battre les humains aux échecs ou au Go c'est une chose. Trouver un vaccin anti-Corona une autre, un peu moins predictable. Dans un ou deux siècles on peut se reposer la question. Mais peut-être serons-nous en monarchie de droit divin ?	2020/09/1
3763613940335005			5d392225c0d98591e13412232b084bff	6730bc966451144407ffc23a7a024971	Lia Yáng Aebi	2020/09/1
3780287218667677			6876d45b18c6cc9ffbeae83dff56e0ad	5aa203075d057b215065c850410ed552	Avec la Chine en 1ère ligne en la matière, c'est mal barré 🤣	2020/09/2
3763964870299912			ea38e9fcafaa99629c15fd7a4ab3b4aa	b39e1f074d53257a834ebd8ed3183bda	Laura Venchiarutti-Tocmacov	2020/09/1
3781925818503817			9900c47ad570be7772a493e25392c804	adcf425d699c4fc194d4181ef4ee6ad4	De toute façon la meilleure démocratie c'est une dictature éclairée.	2020/09/2
3183512455011826			ea38e9fcafaa99629c15fd7a4ab3b4aa	b39e1f074d53257a834ebd8ed3183bda	Laura Venchiarutti-Tocmacov	2020/02/2
3184302798266125			83201c19ca4a1c60b08e90c2c96155cb	5aa9eb1570e430fecef0243edf56b220	Nom de la chaine science4all si ça intéresse des gens ^^	2020/02/2
3184250951604643			006e1bc135e27655dd880ae787e1d8f6	ddf94015662074bfe6ade490b7ae18cd	L'AI de Youtube sert les interets de son proprietaire comme toutes les AI. Son proprietaire ce n'est pas l'humanité mais les investisseurs. Il n'y a aucune raison pour qu'elle maximize le bonheur. Elle obeit a son maitre et maximize donc le profit.	2020/02/2

Downstream: Analytical Challenges

- Facebook comments tend to mimick spoken language, and are subject to a very wide topical variation.
- We have been bad so far at implementing automatic topic modelling.
  - LDA techniques has performed unsatisfactorily.
  - Named entity recognition algorithms also performed badly.
  - “Basic” NLP techniques (e.g. analysis of frequencies of n-grams) have returned more solid results.
  - A lot of manual analysis & coding has been involved.
- Visual communication is not, at this stage, captured by the analysis (memes etc.)



Activity Patterns



REPUBBLICA TOTAL POSTS

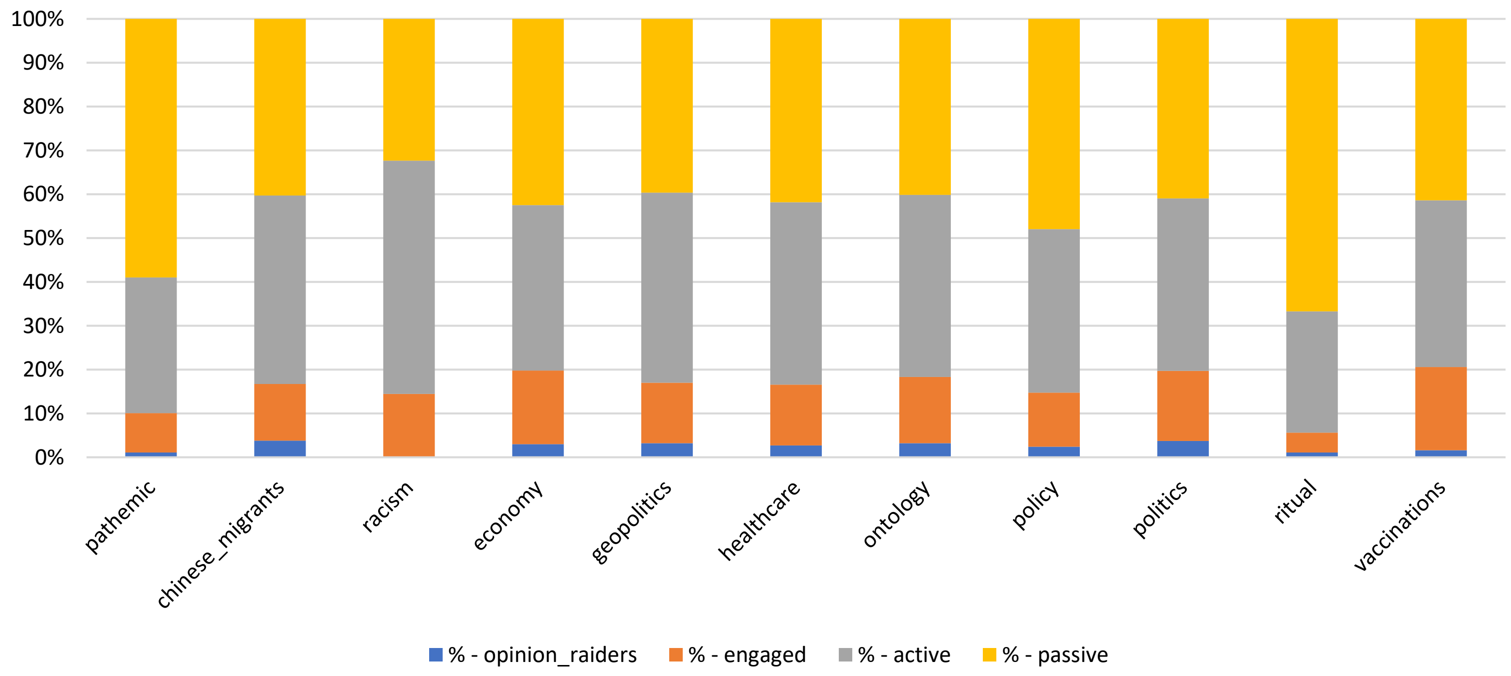
Virus as a Chinese Problem

Covid in Italy

Lockdown

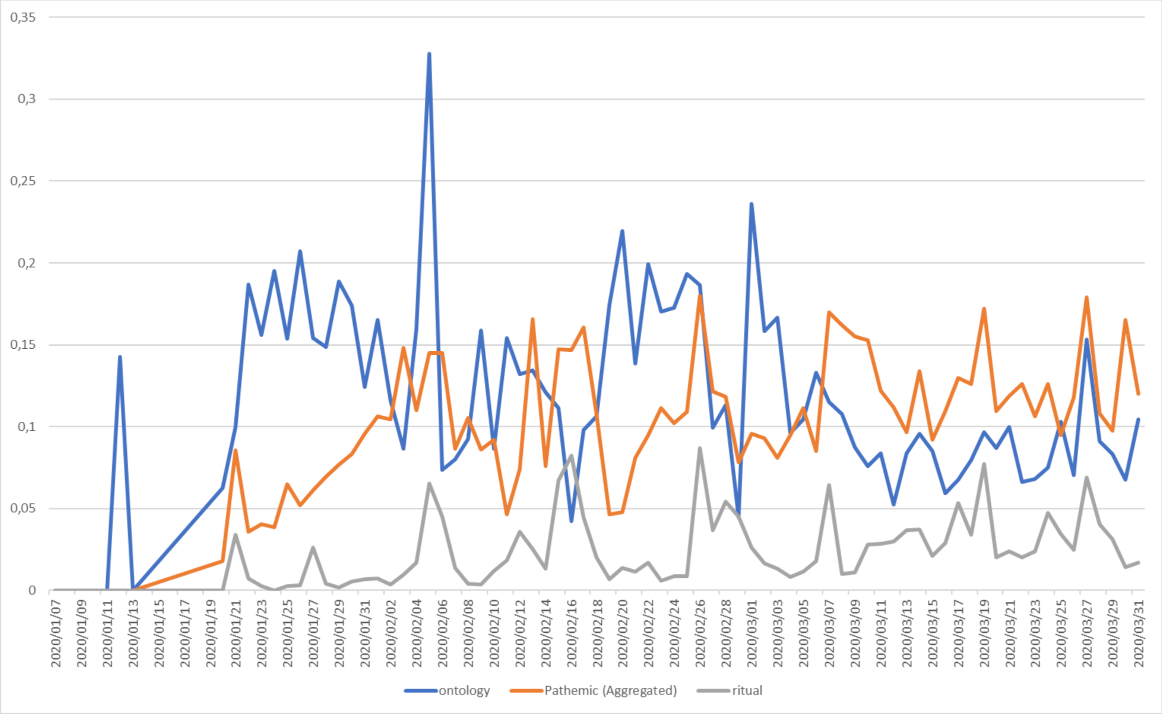
User profiling

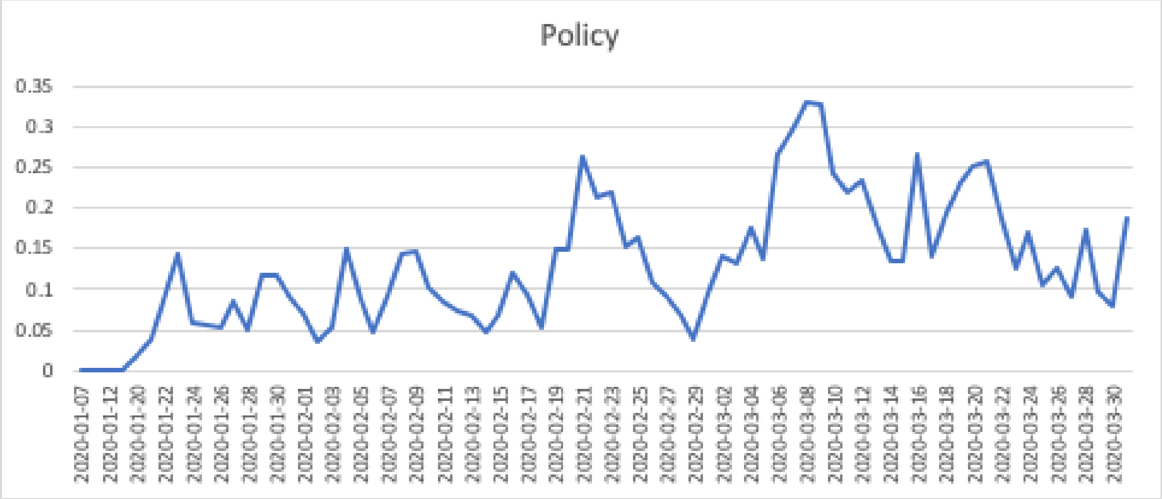
PROFILE	AVG_LENGTH	TOT_COMMENTS	AVG_REPLIES	LEXICAL_VARIETY
OPINION RAIDERS	10.56378936	11052	1.921001	0.456169
ENGAGED	9.846036024	67899	1.207014	0.666132
ACTIVE	8.505942023	235610	0.967637	0.853186
PASSIVE	6.34536352	376719	0.508742	0.936105



Topics By User Profile

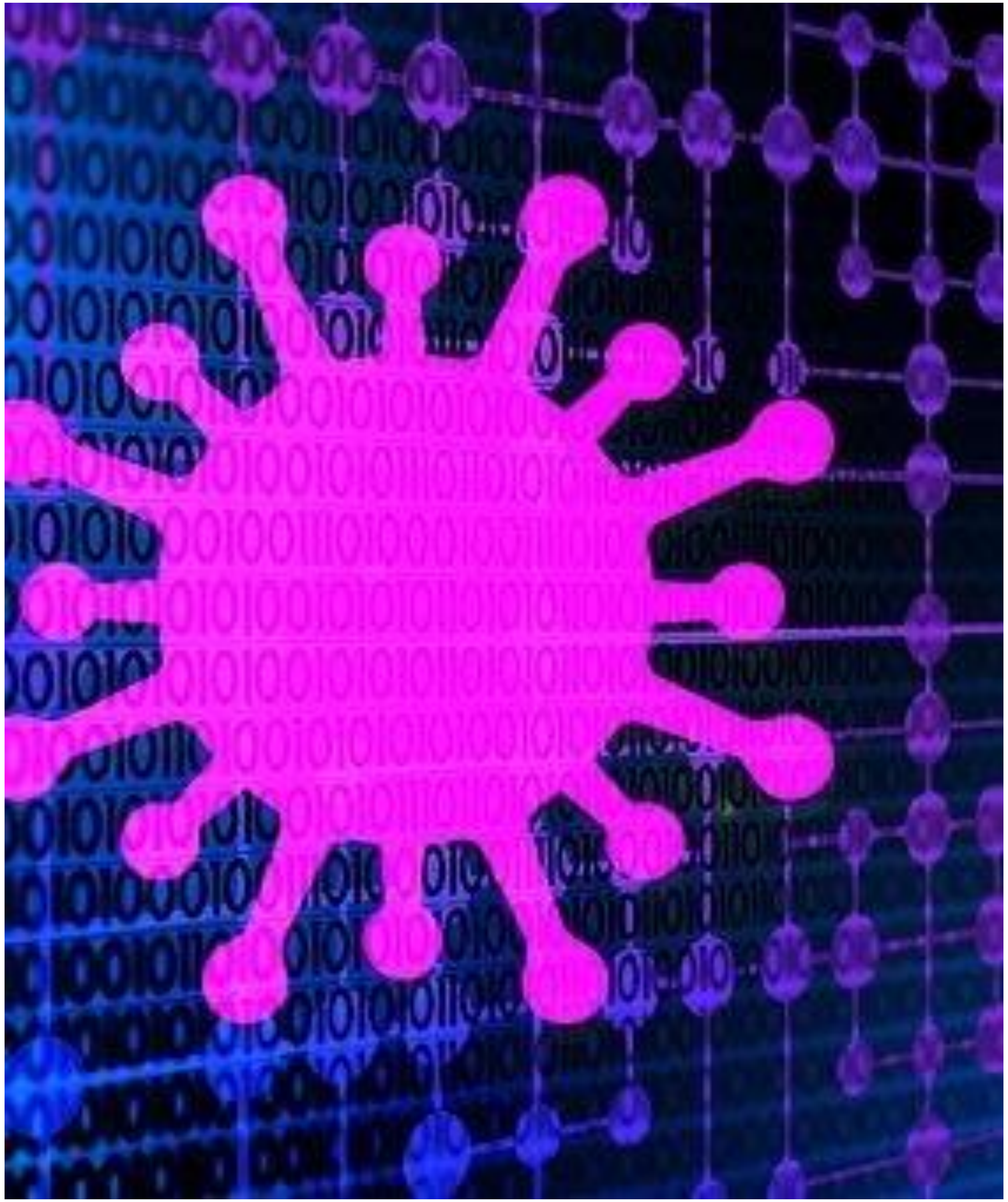
Topical Trends





Next Steps

- Applying those techniques to understand how AI is thematized in Facebook through the analysis of five newspapers (Guardian, Le Temps, LeMonde, Independent, New York Times)
- Degree of uncertainty about Facebook «next moves»; difficult to plan ahead.



Conclusions

- Understanding Facebook discourse dynamics is a core concern, particularly in times of crisis.
- Facebook terms of use appear **way too restrictive;** solid, empirically-grounded research becomes very expensive iunder them.
- But even if we got Facebook’s permission to harvest its data, its limitations and constraints still make data collection and processing very costly and risky.
- The academic community should fight for easier academic access to Facebook data, while preserving privacy.