Beyond twitter

Exploring `bluesky.social` for digital disease detection and prototyping a data extraction pipeline for ILI surveillance

Heiner Atze, MSc, PhD

Digital Epidemiology 2025, Hasselt University

2025-04-10

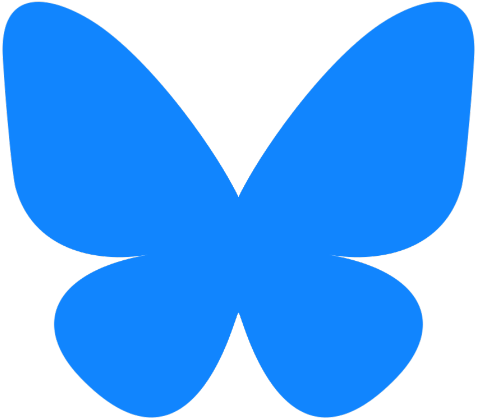
Outliningglines I

1. Introduction
2. Exploration of bluesky data
3. Project
4. Methods
5. Data extraction
6. Results

Introduction

bluesky: general aspects

- microblogging platform
- similar to twitter in user experience
- decentralized
- open source



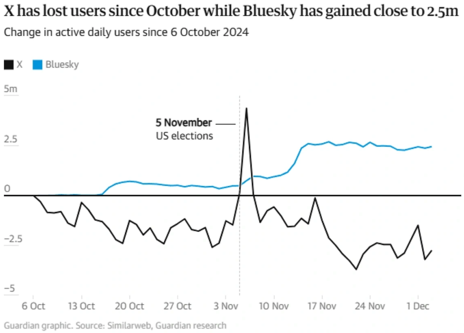
Decentralization and Democratization of content algorithms [1]

- Decentralized User Identifier (DID)
    - immutable, associated with human readable user handle
- Personal Data servers (PSDs)
- DIDs and affiliated contents are portable between PSDs
- Users can choose, prioritize and develop feed generators and content labelers

[1] Balduf et al. (2024)

Development of user activity [2]

- current estimate: ca. 33 Millions active users
- user base expanded in bursts after key events:
  - 2022: acquisition of `twitter` by Elon Musk
  - 2024: ban of `X` in Brazil, presidential election in the US



X has lost users since October while Bluesky has gained close to 2.5m

Change in active daily users since 6 October 2024

Guardian graphic. Source: Similarweb, Guardian research

[2] Duarte, Balduf et al. (2024)

Literature addressing `bluesky`

- Google scholar search : “bluesky” AND “social” since 2022
- 43 articles
- main topics:
  - decentralized social network architecture
  - user migration from `X` to `bluesky` 2024
  - network structure and dynamics
- no results for
  - “bluesky” AND “disease”
  - “bluesky” AND “epidemiology”

Exploration of bluesky data

bluesky API

- publicly accessible for free
- extensive documenation at https://docs.bsky.app/docs/category/http-reference

`searchPosts` API method

- API documentation
- selected parameters:
  - `q`: search query
  - `since`, `until`: defining search period
- deterministic search
- allows exhaustive sampling

getProfiles

- allows to retrieve the author profile information
- for reference, not used in this project

Post metadata

- defined in the SDK documentation
- fields (selection):
  - `uri`: unique post identifier
  - `author`: contains `did` which allows to retrieve user profile
  - `record`: contains the text and time information of the message
  - `embedded`: any embedded media (images, other posts, etc ...)
- in contrary to former `twitter` post metadata, no geoinformation

User information

- `Feedgens`
- `Labelers`
- no geo information

Project

Outline

`bluesky` post data for digital disease surveillance

Outline

`bluesky` post data for digital disease surveillance

Implementation of a continuous surveillance pipeline

Methods

Data extraction

Symptom related message extraction

- focused on French `bluesky` posts (data volume constraint)
- extraction using list of keywords
    - grippe (flu, influenza)
    - rhume (common cold)
    - fievre (fever)
    - courbature (muscle pain)
- extraction of
    - complete message data for further language processing
    -

Basal network activity

- probing of the basal network activity using keywords
    - travail (*work*)
    - demain (*tomorrow*)
    - voiture (*car*)
    - sommeil (*sleep*)
- post counts aggregated by day

Case data

- data downloaded from `WHO Flumart` = FluID: ILI case data
    - FluNet: virological data

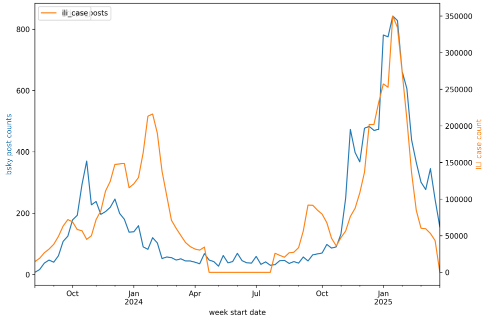
Data processing for time series extraction

- Normalization of ILI post counts by basal network activity
-
- LLM
- ECDC case definition
  - LLM vs. random post selection

Results

Raw post counts

```
Text(0, 0.5, 'ILI case count')
```

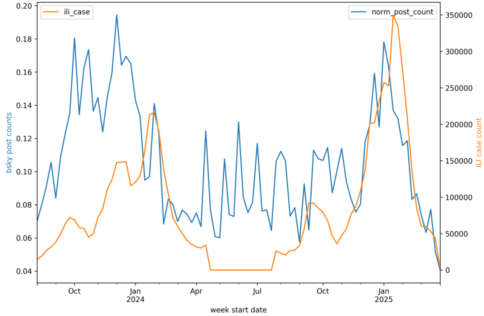


Correlation

	gripppe_posts	rest_posts
gripppe_posts	1.000000	0.878014
rest_posts	0.878014	1.000000
ili_case	0.775933	0.568114

Normalized post counts

```
Text(0, 0.5, 'ILI case count')
```



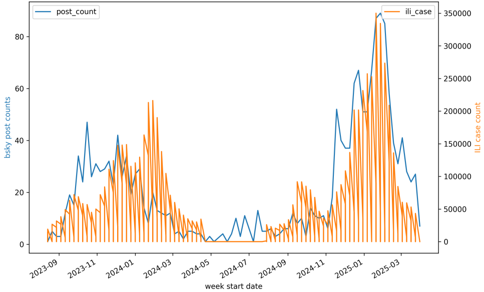
It is not as simple as that .... :/

Correlation

	norm_post_count	re
norm_post_count	1.000	0.
rest_posts	0.063	1.
ili_case	0.515	0.

LLM annotated post counts, raw

```
Text(0, 0.5, 'ILI case count')
```



Correlation

	ili_case	post_count
ili_case	1.000	0.396
post_count	0.396	1.000

Bibliography

Balduf, Leonhard, Saidu Sokoto, Onur Ascigil, Gareth Tyson, Björn Scheuermann, Maciej Korczyński, Ignacio Castro, and Michał Król. 2024. "Looking at the Blue Skies of Bluesky." In *Proceedings of the 2024 ACM on Internet Measurement Conference*, 76–91.

Duarte, Fabio. "Bluesky User Age, Gender, & Demographics (2025)." https://explodingtopics.com/blog/bluesky-users.