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

Outlininglines I

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

Outlininglines II

7. Post count time series

8. Can “AI” help?

Outline

- The `bluesky` social network
- Data accessiblity via the `bluesky` API
- Project: Analysis of ILI related `bluesky` messages

Introduction

bluesky: general aspects

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

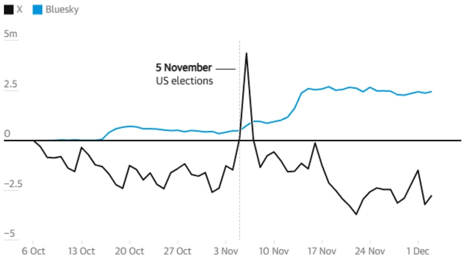
Decentralization and Democratization of content algorithms [1]

- 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

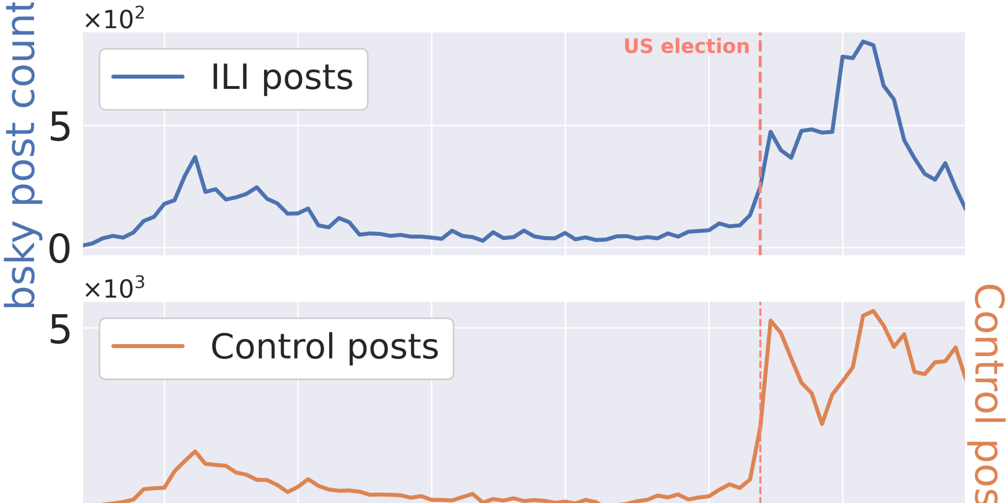
- Normalization of ILI post counts by basal network activity
-
- LLM
- ECDC case definition
  - LLM vs. random post selection

Results

Post count time series

Raw posts counts

- Data analysis starting from 2023-08-01



Keyword posts *vs.* ILI incidence

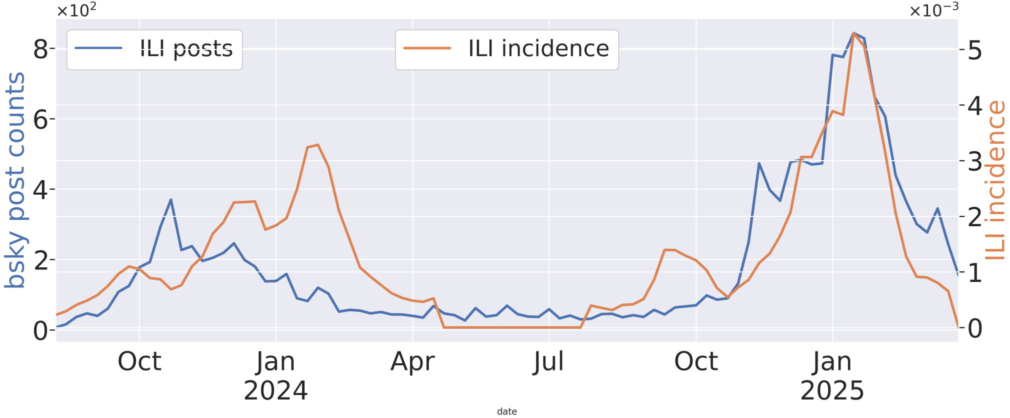


Figure 2

	ILI posts	Control posts	ILI cases
ILI posts	1.000	0.878	0.775

Normalized keyword posts *vs.* ILI incidence

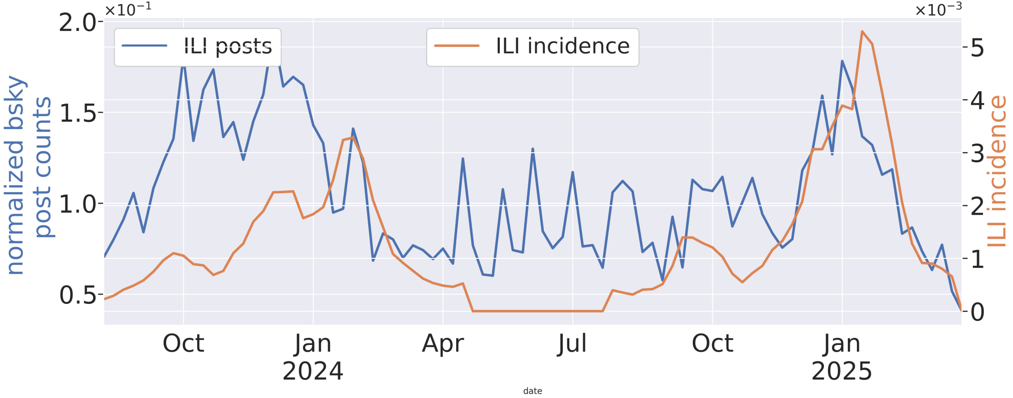


Figure 3

- Normalization of the number ILI keyword containing messages using the number of control messages

	ILI posts	Control posts
ILI posts	1.000	0.063

Machine Learning - Features

- no. of control posts
- no. of posts containing ILI related keyword
- seasonal features: year, month, week, season
- lag terms

Machine Learning - Gradient boosted trees

- Sequential learning of weak learners.
- Iteratively corrects errors of previous models
- Combines predictions using weighted averaging.
- Robust to outliers due to tree-based structure.
- Handles non-linear relationships

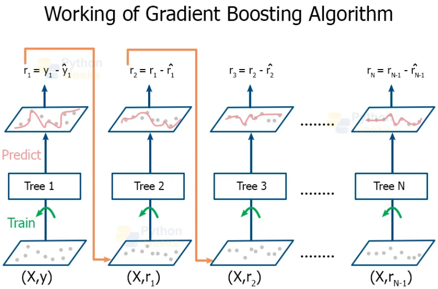


Figure 4: Gradient boosting [a]

[a] Team

Machine Learning - Model evaluation

- Time series split validation
  - retains temporal information
  - mimics continious data acquisition

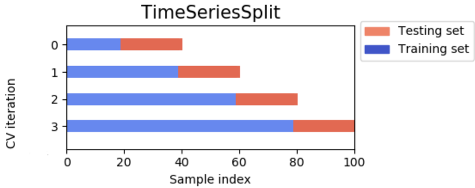
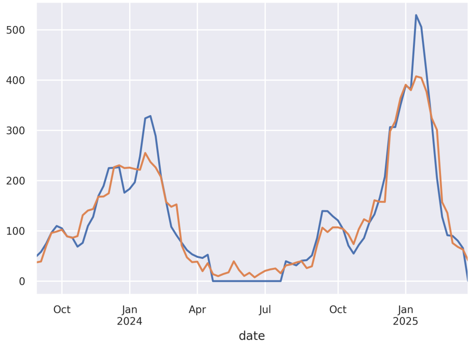


Figure 5: Expanding window time series validation [a]

[a] "How to Apply Stacking Cross Validation for Time-Series Data? — Datascience.stackexchange.com"

Machine Learning = results

- Target variable: ILI incidence one week ahead
- small sample size: Generalization error approximation by validation



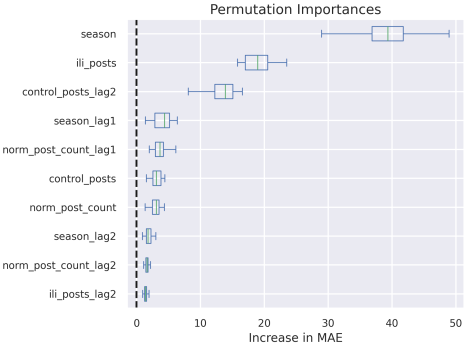
Metrics

Dataset	MAE*
Training	23.79
Validation	80.61

* Mean absolute error, incidence per 100,000

Permutation importance

- model agnostic feature importance procedure
- random shuffling of single input features



Can “AI” help?

Idea

- Filter posts using large a large language model (LLM)

How?

- provide case definition in the system prompt
- use `json` structured output option for convenient data processing

Prompt and output

Prompt extract

```
Analyze the following tweet-like m                                es

- Fever  38°C (100°F) **AND**

- At least one respiratory symptom

- Additional systemic symptoms (he                                ig
...
```

```
// symptom extraction schema
{
    "ili_related" :{
        "type":"bool"
    },
    symptoms:{
        "type":"array",
        "items":{
            "type":"string"
        }
    }
}
```

Examples

ILI positive

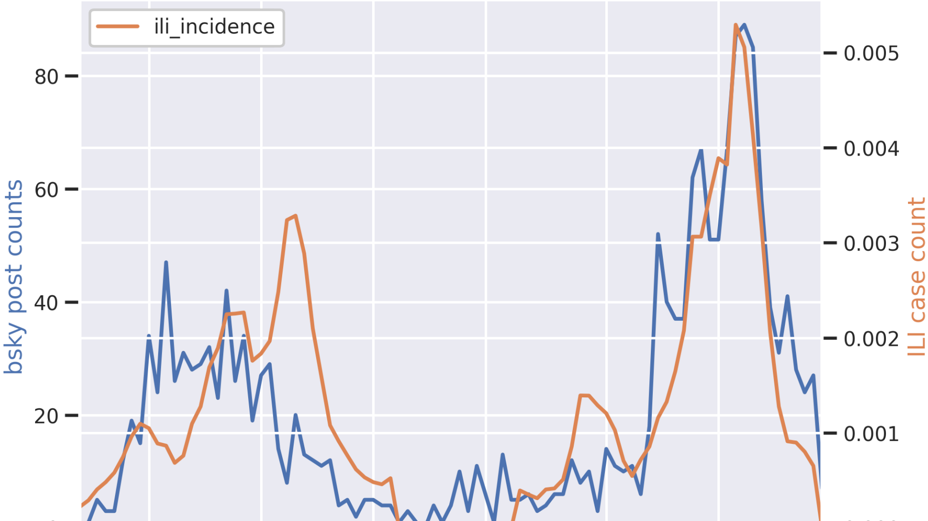
Original message	Machine translation	LLM sym
Attrapé début octobre une semaine avant d'aller me faire vacciner	Caught in early October a week before g	Fièvre, f

ILI negative

Original message	Machine translation	LLM sym
Grippe aviaire : les coupes budgétaires de Trump amplifient	Avian flu: Trump's budget cuts amplify pa	nan

LLM annotated post counts, raw

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



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.

“How to Apply Stacking Cross Validation for Time-Series Data? — Datascience.stackexchange.com.” https://datascience.stackexchange.com/questions/41378/how-to-apply-stacking-cross-validation-for-time-series-data.

Team, PythonGeeks. “Gradient Boosting Algorithm in Machine Learning - Python Geeks — Pythongeeks.org.” https://pythongeeks.org/gradient-boosting-algorithm-in-machine-learning/.