Beyond twitter

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

Heiner Atze

Digital Epidemiology 2025, Hasselt University

2025-04-10

Outlininglines I

Outlininglines II

Outline

Outline

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

Introduction

`bluesky`: general aspects

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

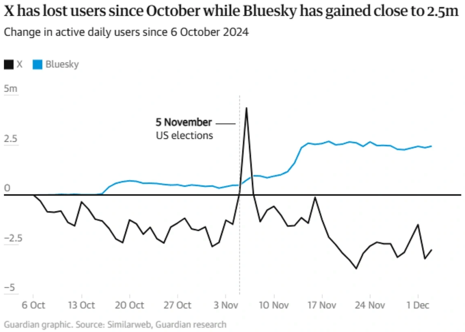
Decentralization and Democratization of content algorithms [1]

- 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



[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
- `limit`: max. 100 posts
- 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
    - `langs`: language(s) detected by the `bluesky` server
  - `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

Data extraction

ILI symptom related message extraction

- focused on French `bluesky` posts (data volume constraint)
- extraction using list of keywords [3]
    - grippe (*flu, influenza*)
    - rhume (*common cold*)
    - fievre (*fever*)
    - courbature (*muscle pain*)
- extraction of
    - complete message data for further language processing
    - counts for time series analysis

[3] Signorini (2011)

Basal network activity

- Keywords:
  - travail (*work*)
  - demain (*tomorrow*)
  - voiture (*car*)
  - sommeil (*sleep*)
- post counts aggregated by day

Case data

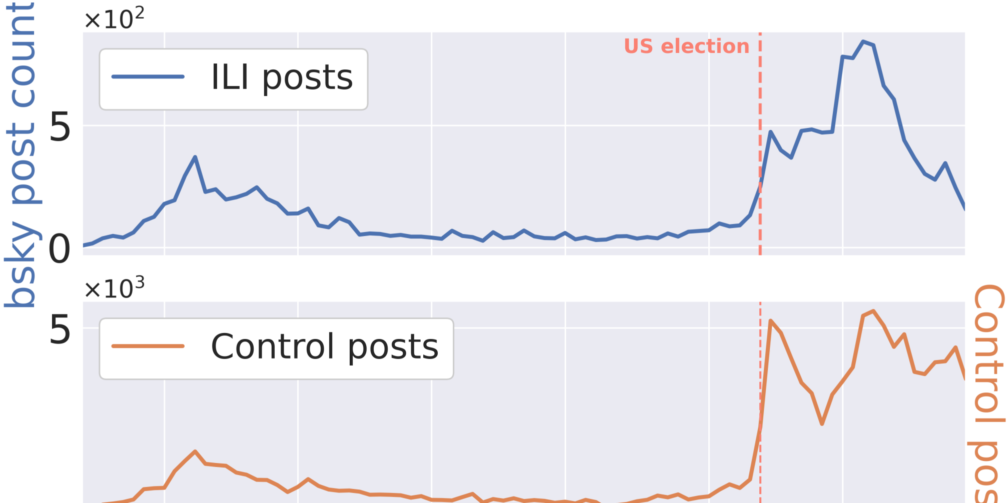
- data downloaded from `WHO Flumart`
  - FluID: ILI case data

Results

Post count time series

Raw posts counts

Data analysis starting from 2023-08-01



Keyword posts *vs.* ILI incidence

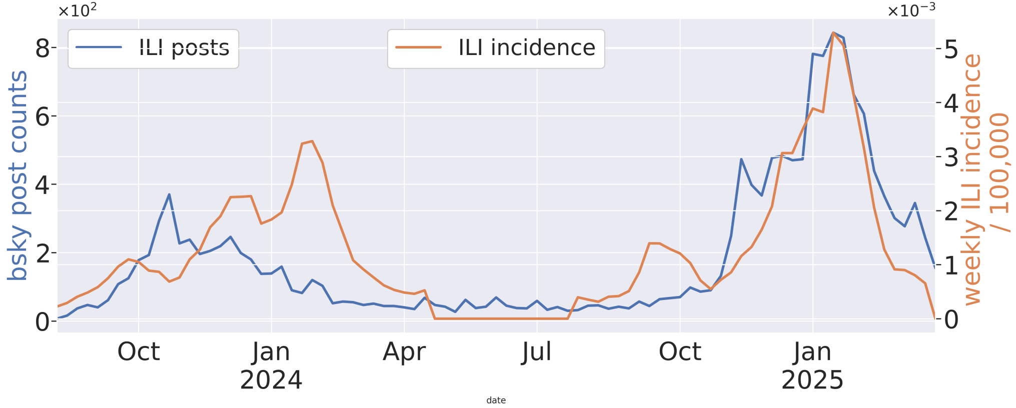


Figure 2

	ILI posts	Control posts	ILI incidence
ILI posts	1.000	0.878	0.775

Normalized keyword posts *vs.* ILI incidence

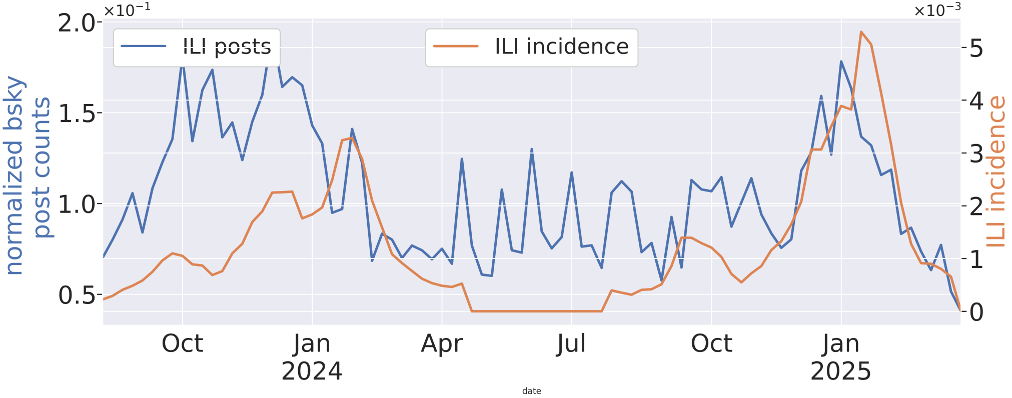


Figure 3

- 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
- time and seasonal features
  - year
  - month
  - week
  - season
- lag terms

all aggregated by week

Gradient boosted trees

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

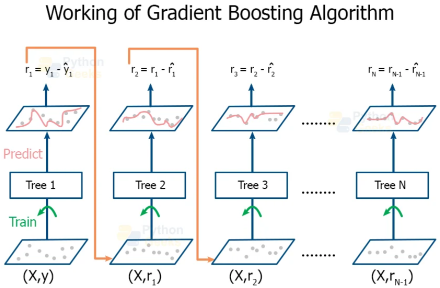


Figure 4: Gradient boosting [a]

[a] Team

Model evaluation

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

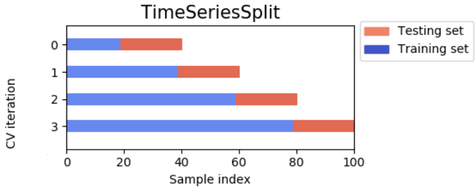


Figure 5: Expanding window time series validation [a]

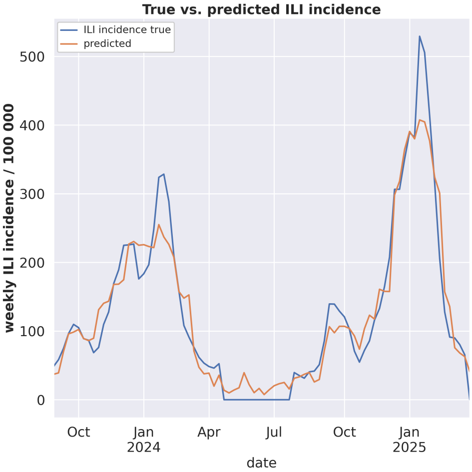
[a] "How to Apply Stacking Cross Validation for Time-Series Data? — Datascience.stackexchange.com"

Predictions and metrics

- Target variable: weeky ILI incidence w_{t+1}

Predictions and metrics

- Target variable: weeky ILI incidence w_{t+1}



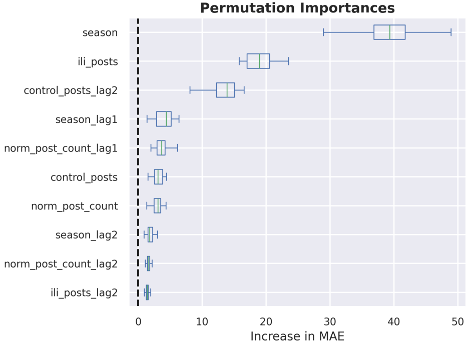
Metrics

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

...


{ ... bluesky message dynamically
```

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

```
// symptom extraction example
{
    "ili_related" :true,
```

Examples

ILI positive

Original message	Machine translation	LLM sym
Oui, j'ai une fièvre mais pas trop forte. Alors que je suis plus … Mais je suis bien KO quand même.	Yes, I have a fever but not too strong. But I'm well Ko anyway.	Fever

ILI negative

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

LLM annotated post counts

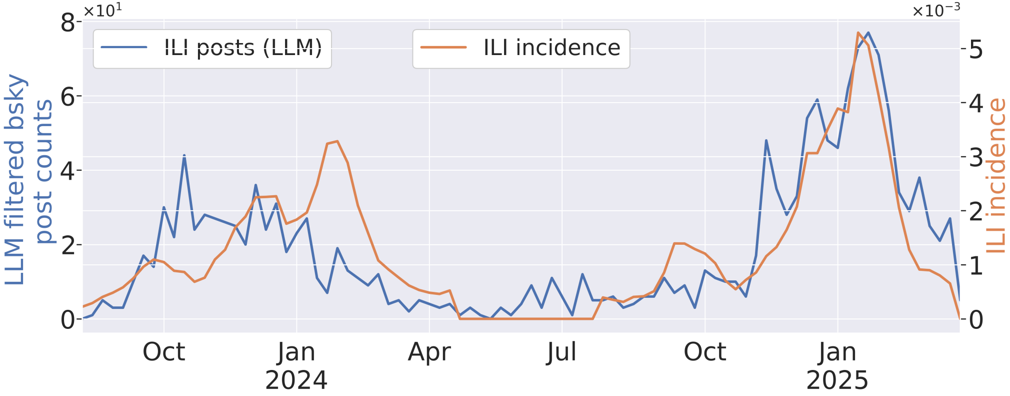


Figure 6

Correlation

	LLM ILI posts	ILI incidence	Control posts

Conclusion

Conclusion

- `bluesky` = promising data source
- more data needed = patience

Outlook

Outlook

- investigate impact of LLM filtering on model performance
- modeling of weekly ILI incidence based on message content
- continuous data acquisition pipeline (WIP)
- User localization based on profile
- monitoring of bursts in user activity crucial
- repeating the analysis for another country (*e.g.* Germany)

Pipeline (WIP)

```
graph LR
    subgraph kestra
        dlt(dlt) --- posts
        llm --- bqstaging
        llm -- annotation --> bqstaging
        posts --> bqstaging[<b>GBQ</b> \n stage area \n 1 table per k
        dlt -- housekeeping --> count
        dlt -- case data --> who_tables
        dlt -- case data --> cdc_tables
        subgraph BigQuery data lake
          bqstaging
          who_tables
          cdc_tables
          count[post counts table]
        end
        bqstaging --- dbt
```

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.

Signorini, Alberto Maria AND Polgreen, Alessio AND Segre. 2011. "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the u.s. During the Influenza a H1N1 Pandemic." *PLOS ONE* 6 (5): 1–10. https://doi.org/10.1371/journal.pone.0019467.