# Mapping Reddit Posts Worldwide

## Final Projekt

290706 LP Capstone Project (2025W)
Mina Karimi, BSc MSc Ph.D.

Presented by Luis Fink, 12018892

# Structure

1. Introduction & Motivation
2. Related Work
3. Research Gap
4. Research Question
5. Data & Methods
6. Results
7. Conclusion

# Introduction & Motivation

- Reddit is one of the world's **largest discussion-based platforms**, with hundreds of millions of monthly active users and millions of discussion-rich and location-oriented communities

- Discussions on Reddit are organised into **topic-specific subreddits**, including country- and location-focused subreddits
  → Perfect for extracting locations and performing spatial analysis

- **Access:** Academic Torrents publishes **complete Reddit datasets dumps** every month for free
  → which enables large-scale spatial analysis with million of posts **without restrictions**

**Motivation:**

- Make global Reddit activity **spatially visible** on a map

- Identify **geographic and socioeconomic biases** in online discussions

- Move beyond raw country mention counts by **comparing county mentions globally** and linking it with country-specific **socio-economic factors**

# Related Work

Geolocation extraction from social media text has been an **active research area for years** , especially for platforms like Twitter (now “X”) or Reddit.

**Examples of related Work:**

- **Harrigian (2018)** conducted one of the first **text-based geolocation analyses of Reddit** posts and users.
  → Showed that location extraction is possible **without explicit GPS tags** by using textual and metadata indicators.

- **Stillman and Kruspe (2024)** performed geolocation experiments on Reddit using **location-focused subreddits and deep learning models** to classify the geographic origin of content at the city level.
  → Their results confirmed that Reddit text contains **plenty extractable spatial information**, but also highlighted **challenges related to data sparsity and noise.**

- **Cheng et al. (2010)** demonstrated that the geographic origin of tweets can be predicted using **location-specific words in tweet text**, even when no explicit geotags are available.
  → Their work showed that tweets alone already contain **strong spatial indicators**.

- **Karami et al. (2021)** analysed Twitter geolocation data and showed that only a **very small and non-representative fraction of tweets is geotagged**.
  → They highlighted how this leads to significant **spatial biases** in social media-based geographic analyses.

# Research Gaps

- Most existing geolocation studies focus mainly on **Twitter** and rely on **limited or biased geotagged data**

- Reddit has been explored for geolocation, but mostly at a small scale (city level) or for specific case studies
  → Large-scale, global Reddit mapping still remains **largely unexplored** with only a few studies addressing scalability to hundreds of millions of posts

- Prior work rarely connects social media attention with **socio-economic factors and country development indicators** such as population size, internet usage or the Human development Index
  → which limits insights and leaving the underlying causes of observed spatial patterns unexplained.

- Most studies focus primarily on **English country and city names**, while multilingual location extraction is frequently overlooked
  → This results in a **strong bias toward English-speaking countries** and limits the global comparability of spatial analyses

# Research Questions

1. How are country mentions on Reddit posts **spatially distributed** across the globe?
2. Which **spatial biases** become visible on a map of global Reddit country mentions?
3. Which countries receive **disproportionately high or low attention** relative to their population size and internet usage rate?
4. How does global Reddit activity relate to a **countries overall development level**, as measured by the United Nations' official Human Development Index (HDI)?

# Data

Reddit Posts were obtained from **Academic Torrents**
*(Source: https://academictorrents.com/browse.php?search=reddit)*

- **Total dataset size:** over 400 million Reddit posts (Compressed file size: ~166.5 GB)
- **Time period:** January to October 2025 (10 months)
- **Data format:** compressed JSON lines (.zst)
- **Extracted textual information:**
  - post titles
  - post bodies
  - subreddit names

  → comments of each post were excluded to reduce the data size

```
Counting: RS_2025-01.zst
Counting: RS_2025-02.zst
Counting: RS_2025-03.zst
Counting: RS_2025-04.zst
Counting: RS_2025-05.zst
Counting: RS_2025-06.zst
Counting: RS_2025-07.zst
Counting: RS_2025-08.zst
Counting: RS_2025-09.zst
Counting: RS_2025-10.zst

==============================
Total posts across all files: 408,837,380
==============================
```

**Linked Data Sources:**

- **World Population**
  *(Source: https://worldpopulationreview.com/countries)*
- **Internet usage per country**
  *(Source: https://ourworldindata.org/grapher/share-of-individuals-using-the-internet?tab=table)*
- **United Nations Human Development Index (HDI) per country**
  *(Source: https://hdr.undp.org/data-center/documentation-and-downloads)*

# Methods

- First downloaded the Monthly Reddit submission files via a **BitTorrent-Client** and loaded them into **JupiterLab**

- To reduce the English-language bias, I built a **multilingual dictionary** of all country names in the **30 major world languages** using the Wikidata-API → covers over **90% of the global speaking population**

- Also added a dedicated list of **popular short country synonyms** to the country matching (e.g. “us” → United States, “uk” → United Kingdom)

- **Main Step:** Extract country mentions of post-titles, -bodies and -subreddits
  (using the multilingual dictionary and the list of short country synonyms)

  - Tested several text extraction methods, but most were **far too slow** for hundreds of millions of posts (would have taken several days to process)

  - The one exception was the **Aho–Corasick algorithm**, which matches thousands of country names simultaneously in a single pass through the text
    → enables a fast and scalable processing of very large datasets

```
!pip install pyahocorasick
```

# Methods

- To account for false positive country matches (especially in short country synonyms like “uk” or “us”)
  → Applied **context-based rules** alongside the Aho–Corasick algorithm
  → so matching terms were only accepted when used in a **clear geographic context** like *“in”, “from” or “to”*.

```
context_patterns = [
    r"\bin\s+{}",
    r"\bfrom\s+{}",
    r"\bto\s+{}",
    r"\bat\s+{}",
    r"\bnear\s+{}",
    r"\baround\s+{}",
    r"\bacross\s+{}",
    r"\binside\s+{}",
    r"\bwithin\s+{}",
    r"\boutside\s+{}",
    r"\bback in\s+{}",
    r"\bdown in\s+{}",
]
```

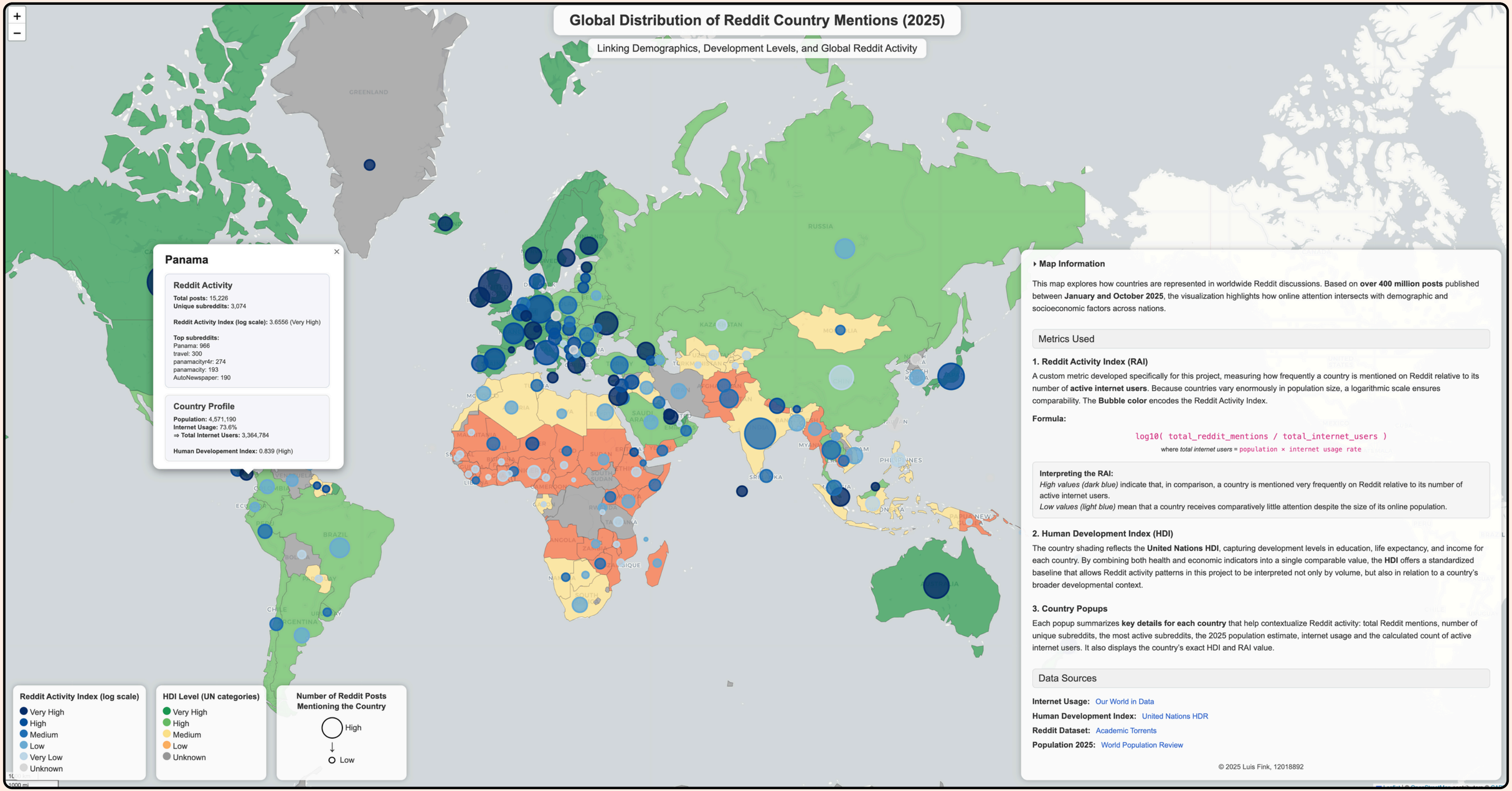
- Developed a custom **Reddit Activity Index (RAI)**
  → Measures how frequently a country is mentioned on Reddit **relative to its total number of active internet users**.
  → Applied a **logarithmic scale** to ensure comparability between countries of totally different sizes

```
log10( total_reddit_mentions / total_internet_users )
```

where *total internet users* = `population × internet usage rate`

- Last Step: **Creating the final Map in Folium**
  - Calculated and aggregated all Matched posts to their corresponding country
  - Linked population size, internet usage and the HDI & RAI to each country
  - Build the Map overlay with Titel, Legends, Popups and a Map information Panel

# Final Map


Global Distribution of Reddit Country Mentions (2025)
Linking Demographics, Development Levels, and Global Reddit Activity
Panama
Reddit Activity
Total posts: 15,226
Unique subreddits: 3,074
Reddit Activity Index (log scale): 3.6556 (Very High)
Top subreddits:
Panama: 966
travel: 300
panamacityr4r: 274
panamacity: 193
AutoNewspaper: 190
Country Profile
Population: 4,571,190
Internet Usage: 73.6%
⇒ Total Internet Users: 3,364,784
Human Developement Index: 0.839 (High)
Map Information
This map explores how countries are represented in worldwide Reddit discussions. Based on over 400 million posts published between January and October 2025, the visualization highlights how online attention intersects with demographic and socioeconomic factors across nations.
Metrics Used
1. Reddit Activity Index (RAI)
A custom metric developed specifically for this project, measuring how frequently a country is mentioned on Reddit relative to its number of active internet users. Because countries vary enormously in population size, a logarithmic scale ensures comparability. The Bubble color encodes the Reddit Activity Index.
Formula:
log10( total_reddit_mentions / total_internet_users )
where total internet users = population × internet usage rate
Interpreting the RAI:
High values (dark blue) indicate that, in comparison, a country is mentioned very frequently on Reddit relative to its number of active internet users.
Low values (light blue) mean that a country receives comparatively little attention despite the size of its online population.
2. Human Development Index (HDI)
The country shading reflects the United Nations HDI, capturing development levels in education, life expectancy, and income for each country. By combining both health and economic indicators into a single comparable value, the HDI offers a standardized baseline that allows Reddit activity patterns in this project to be interpreted not only by volume, but also in relation to a country's broader developmental context.
3. Country Popups
Each popup summarizes key details for each country that help contextualize Reddit activity: total Reddit mentions, number of unique subreddits, the most active subreddits, the 2025 population estimate, internet usage and the calculated count of active internet users. It also displays the country's exact HDI and RAI value.
Data Sources
Internet Usage: Our World in Data
Human Development Index: United Nations HDR
Reddit Dataset: Academic Torrents
Population 2025: World Population Review
© 2025 Luis Fink, 12018892
Reddit Activity Index (log scale)
Very High
High
Medium
Low
Very Low
Unknown
HDI Level (UN categories)
Very High
High
Medium
Low
Unknown
Number of Reddit Posts Mentioning the Country
High
Low

# Results

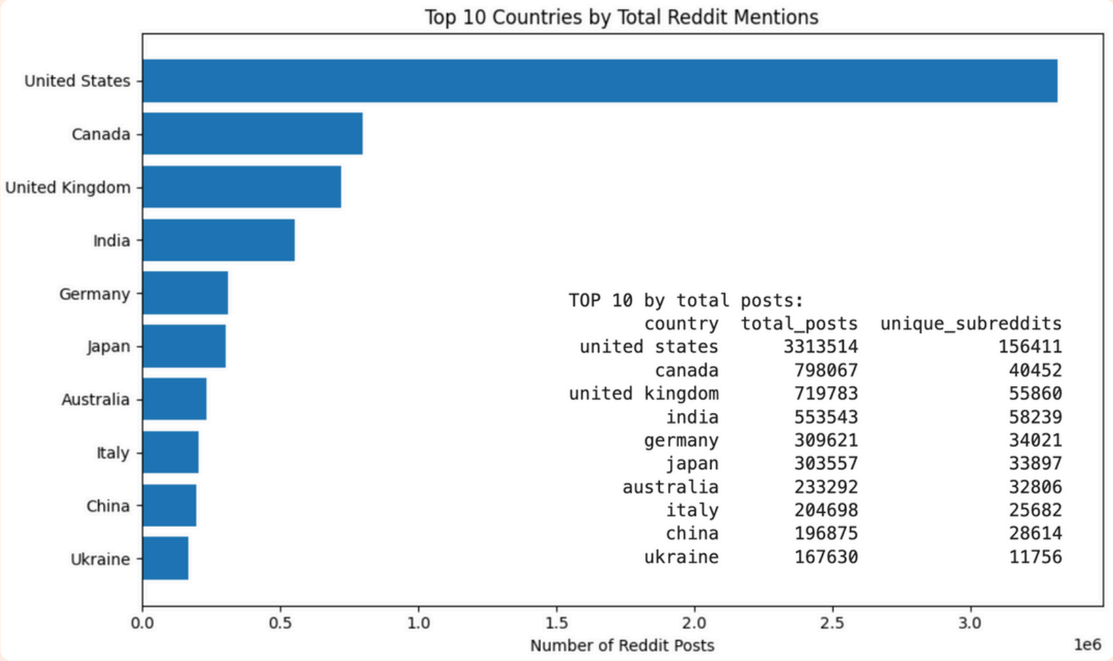
186 different countries were mentioned across 408 million reddit posts in the 30 major world languages

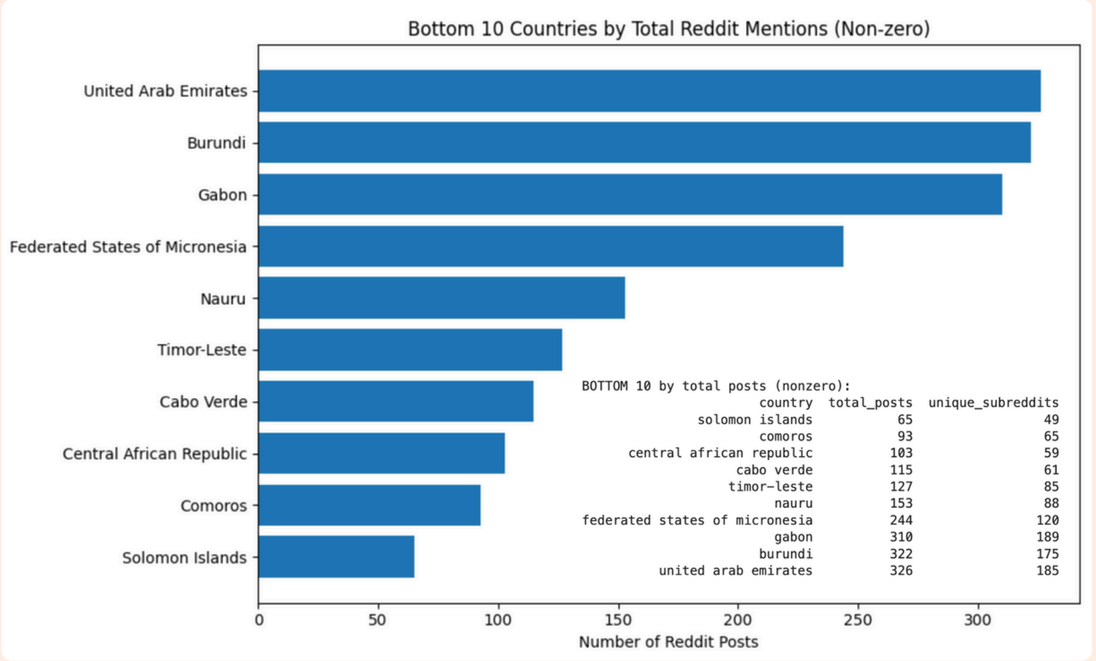
- In total, ~9.3 million Reddit posts were matched to countries (~2.3 % of all posts)
- and a total of ~1.1 million unique subreddits

**Total Post Count**:

→ **United States** dominates with over **3.3 million mentions**, followed by **Canada (~800k)** and **U.K. (~720k)**

→ **Median** country mentions is at only **~7,000 mentions** with many countries receiving fewer than 1,000 mentions



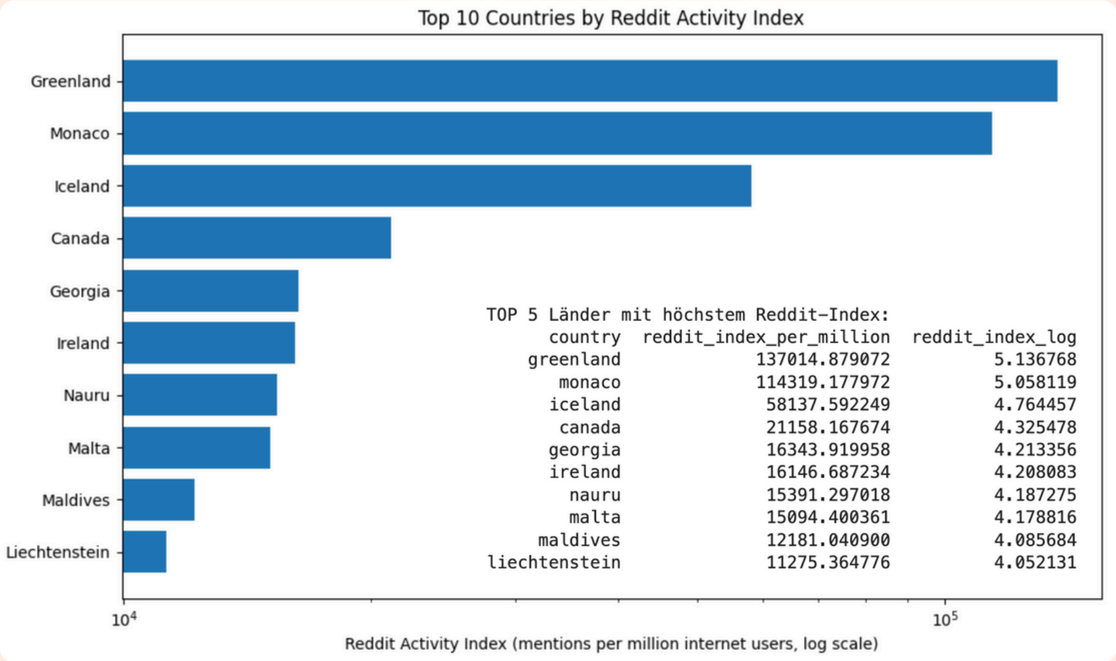
# Results

**Reddit Activity Index (RAI)**: (Country mentions relative to the total number of active internet users)

→ RAI reveals **strongly different patterns** than raw total post counts per country
(Greenland: ~137,000 mentions per million internet users; United Arab Emirates: <30 mentions p.m.i.u.)

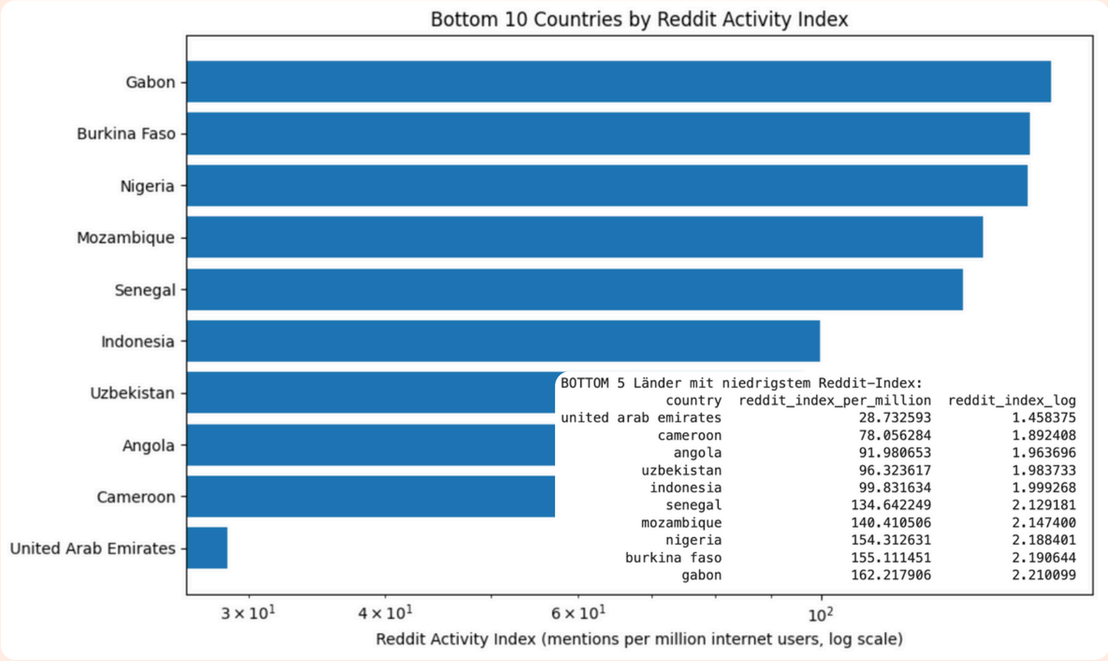
→ **Highest RAI** values are mostly observed in **small, highly concentrated countries** like Monaco, Iceland, Malta or Lichtenstein

→ Countries with **large populations or low development levels** often show low relativ attention




TOP 5 Länder mit höchstem Reddit-Index:

| country | reddit_index_per_million | reddit_index_log |
|---|---|---|
| greenland | 137014.879072 | 5.136768 |
| monaco | 114319.177972 | 5.058119 |
| iceland | 58137.592249 | 4.764457 |
| canada | 21158.167674 | 4.325478 |
| georgia | 16343.919958 | 4.213356 |
| ireland | 16146.687234 | 4.208083 |
| nauru | 15391.297018 | 4.187275 |
| malta | 15094.400361 | 4.178816 |
| maldives | 12181.040900 | 4.085684 |
| liechtenstein | 11275.364776 | 4.052131 |




BOTTOM 5 Länder mit niedrigstem Reddit-Index:

| country | reddit_index_per_million | reddit_index_log |
|---|---|---|
| united arab emirates | 28.732593 | 1.458375 |
| cameroon | 78.056284 | 1.892408 |
| angola | 91.980653 | 1.963696 |
| uzbekistan | 96.323617 | 1.983733 |
| indonesia | 99.831634 | 1.999268 |
| senegal | 134.642249 | 2.129181 |
| mozambique | 140.410506 | 2.147400 |
| nigeria | 154.312631 | 2.188401 |
| burkina faso | 155.111451 | 2.190644 |
| gabon | 162.217906 | 2.210099 |

# Results

**Reddit Activity Index in correlation with Human Development Index:**

- Reddit Activity Index shows a moderate positive correlation with HDI

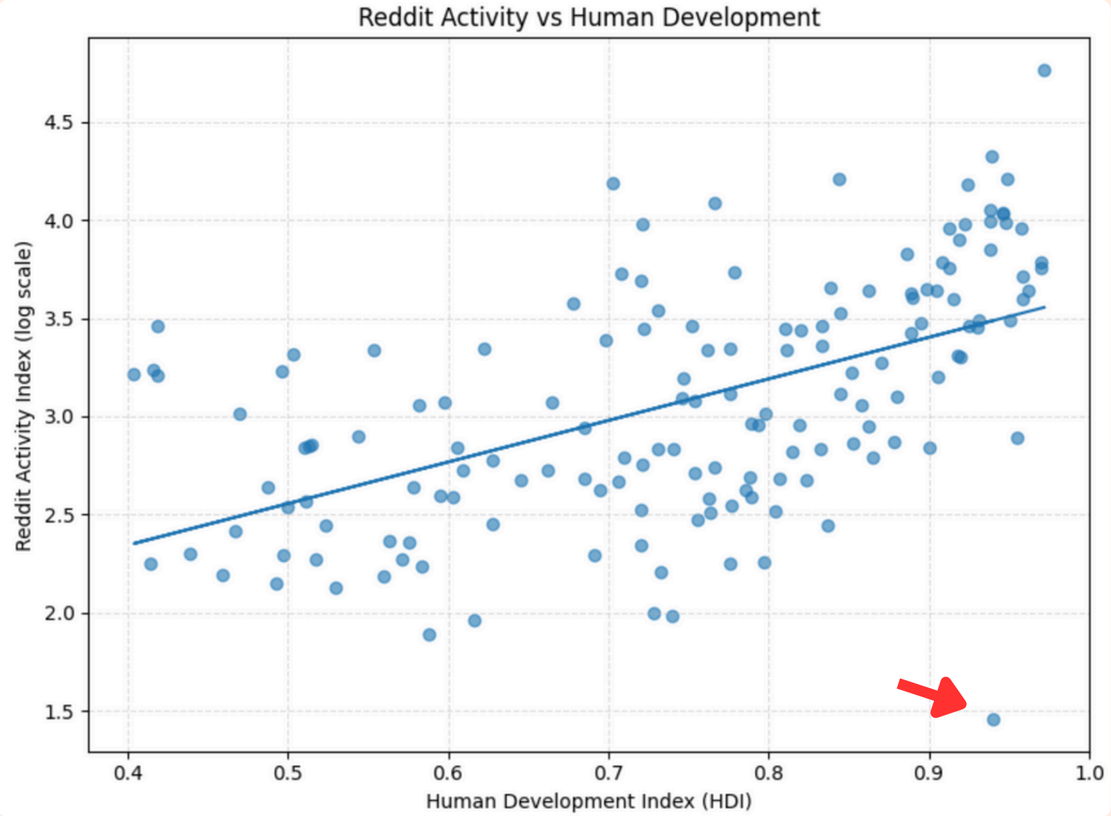
```
Correlation RAI (log) vs HDI
Pearson: 0.544
Spearman: 0.594
```

- Median RAI increases systematically with HDI level

**=> Trend-line Shows:**
Countries with a high Reddit Activity Index,
also tend to have higher human development levels.

**Notable outlier:** The United Arab Emirates combines a very high HDI with the lowest Reddit Activity Index

→ Possibly due to cultural factors or the use of alternative social media platforms



# Results

- **Overall:** Reddit attention is **highly concentrated**

→ **Top 5** countries accounting for over 60% of all mentions
→ **Top 20** countries accounting for over 85% of all mentions

```
Share of total mentions by TOP 5: 60.96%
Share of total mentions by TOP 10: 72.80%
Share of total mentions by TOP 20: 81.58%
Share of total mentions by TOP 30: 86.81%
```

**Final important Notes:**

- Whenever small countries dominate the top ranks in Reddit mentions
  → It often reflects a highly discussed event happening there or other thematic attention
  - Examples:
    - **Ukraine**, which appears in the top 10 in total country Reddit mentions due to the ongoing war leading to wide-ranging global discussions.
    - **Greenland**, which ranks the highest in the Reddit Activity Index (RAI), possible driven by statements by the U.S. President Donald Trump about U.S. interest in the territory, which gained traction in early 2025

- Reddit is mainly **Western and English-speaking**
  → which causes Western countries to dominate the results even with the multilingual dictionary included.

# Conclusion

- Global Reddit attention is **highly uneven** and **dominated by a small number of countries**
- There is a **strong contrast** between **absolute counts of country mentions** and **country mentions relative to a it's online population (RAI)**.
- Reddit activity Index (RAI) of a country generally **increases** with its human development level
- **Small countries** can gain high attention due to **event-driven global discussions** like the war in Ukraine

**Take Home Message:**

→ Overall, this project showed that Reddit Activity patterns are not evenly distributed around the globe, but shaped by **language**, **certain global events** and **broader socio-economic conditions of a country**.

# Sources


- Harrigian, K. (2018). Geocoding Without Geotags: A Text-based Approach for reddit. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1810.03067
- Stillman, M., & Kruspe, A. M. (2024). Geolocation Extraction From Reddit Text Data. In GeoExT@ ECIR (pp. 11-20).
- Cheng, Z., Caverlee, J., & Lee, K. (2010). You Are Where You Tweet: A Content-Based Approach to Geo-Locating Twitter Users. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10).
- Karami, A., Kadari, R. R., Panati, L., Nooli, S. P., Bheemreddy, H., & Bozorgi, B. (2021). Analysis of Geotagging Behavior: Do Geotagged Users Represent the Twitter Population? ISPRS International Journal of Geo-Information, 10(6), 373. DOI:10.3390/ijgi10060373.
- Search for "Reddit" - Academic Torrents. (n.d.). Academic Torrents. https://academictorrents.com/browse.php?search=reddit
- Share of the population using the Internet. (n.d.). Our World in Data. https://ourworldindata.org/grapher/share-of-individuals-using-the-internet?tab=table
- United Nations. (n.d.). Documentation and downloads. Human Development Reports. https://hdr.undp.org/data-center/documentation-and-downloads
- Total population by country 2026. (2026, January 19). World Population Review. https://worldpopulationreview.com/countries

# THANK YOU FOR LISTENING!