Multilingual narrative tracking in the news - real-time experiments

Unstructured Data Meetup Berlin
2024-05-07

Robert Caulk, PhD & Elin Törnquist, PhD

CEO *Director of Transparency*

Team:

Timothy Pogue, Wagner Costa Santos, Emre Suzen

Emergent Methods

Our background

Engineers and Researchers committed to FOSS

- **Applying** AI to real-time adaptive modeling challenges
- **Scaling** software in all directions
- **Enriching** data for other businesses
- **Performing** research

AskNews

News context engineering
https://asknews.app

flowdapt

FOSS 🤗

Real-time cluster orchestration
https://flowdapt.ai

MELISSA

FOSS 🤗

Large-scale deep-learning for supercomputers
https://melissa.gitlabpages.inria.fr/melissa/

FreqAI

FOSS 🤗

AI/ML for algo-trading
https://www.freqtrade.io/en/stable/freqai/

Manifest

FOSS 🤗

Modern configuration
https://github.com/emergentmethods/python-manifest

DATASIEVE

FOSS 🤗

Data pipelining
https://github.com/emergentmethods/datasieve

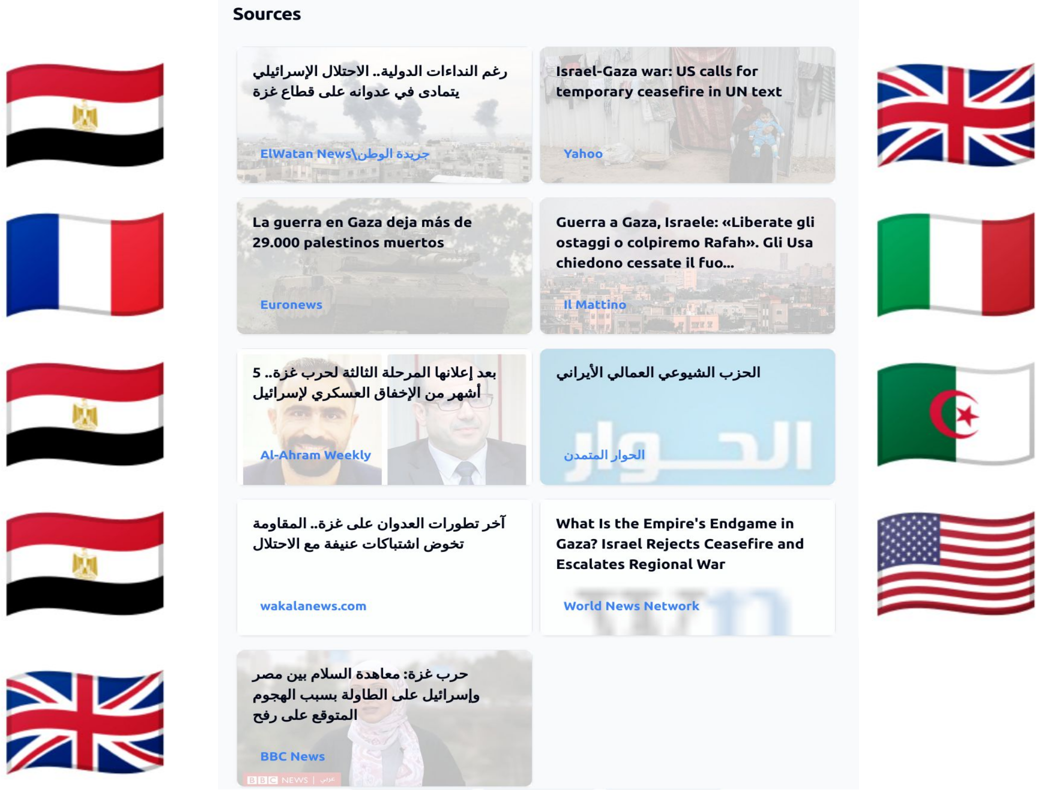
2024-05-07 Unstructured Data Meetup: Multilingual narrative tracking in the news - real-time experiments

Why do we need to engineer news context?

Emergent Methods

Motivations for engineering news context

- Enforcing journalistic standards 📰
  - Stating claims with supporting **evidence** and **attribution**
  - AP style-guidelines and formatting
- Enforcing source and language diversity for democratized representation 🇪🇺
  - Representing **diverse perspectives** on global issues



Sources

رغم النداءات الدولية.. الاحتلال الإسرائيلي يتمادى في عدوانه على قطاع غزة
ElWatan News\جريدة الوطن

Israel-Gaza war: US calls for temporary ceasefire in UN text
Yahoo

La guerra en Gaza deja más de 29.000 palestinos muertos
Euronews

Guerra a Gaza, Israele: «Liberate gli ostaggi o colpiremo Rafah». Gli Usa chiedono cessate il fuo...
Il Mattino

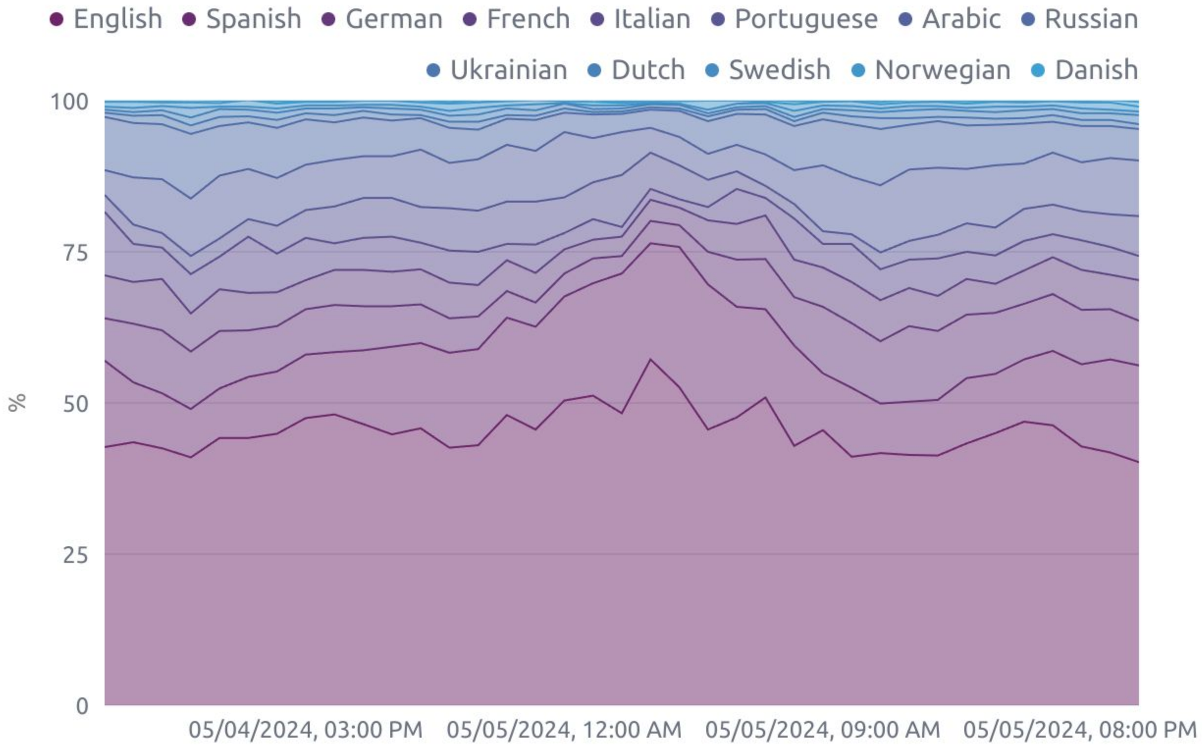
بعد إعلانها المرحلة الثالثة لحرب غزة.. 5 أشهر من الإخفاق العسكري لإسرائيل
Al-Ahram Weekly

الحزب الشيوعي العمالي الأيراني
الحوار المتمدن

آخر تطورات العدوان على غزة.. المقاومة تخوض اشتباكات عنيفة مع الاحتلال
wakalanews.com

What Is the Empire's Endgame in Gaza? Israel Rejects Ceasefire and Escalates Regional War
World News Network

حرب غزة: معاهدة السلام بين مصر وإسرائيل على الطاولة بسبب الهجوم المتوقع على رفح
BBC News



Motivations for engineering news context

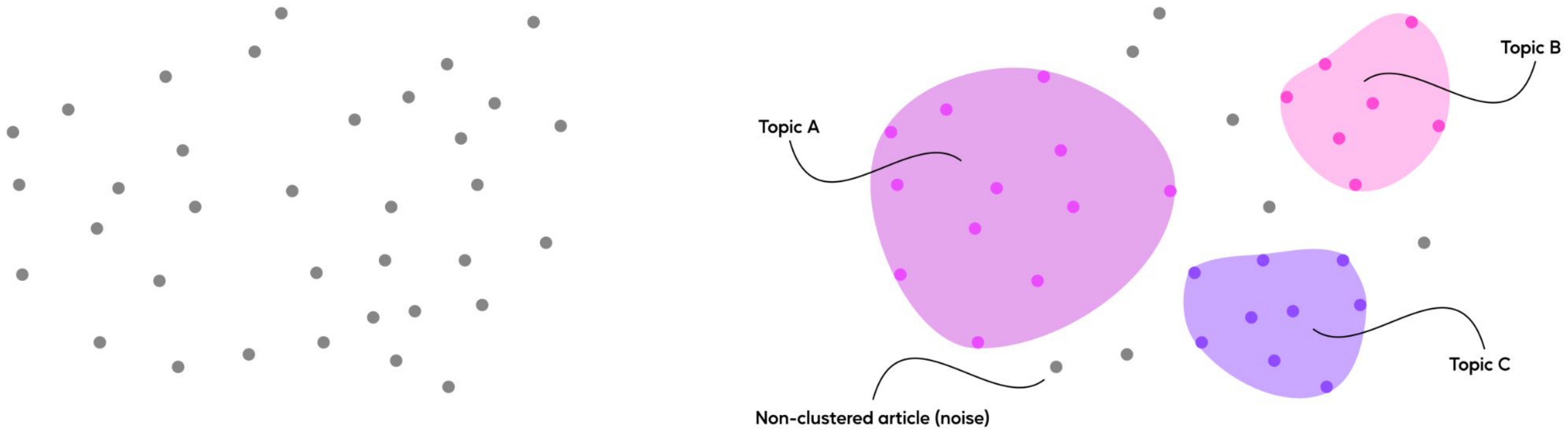
- Enforcing journalistic standards 📰
  - Stating claims with supporting **evidence** and **attribution**
  - AP style-guidelines and formatting
- Enforcing source and language diversity for democratized representation 🇪🇺
  - Representing **diverse perspectives** on global issues
- Avoiding stale/outdated reporting 🏇
  - Missing the latest news can cause **disinformation**, customer dissatisfaction, and **confusion**
- Minimizing hallucination 🫣
  - The cost of hallucination is too high, **a team of researchers is required**
- Scaling democratized news context to companies that are not interested in the logistics of tracking and diversifying 1 million articles per day 🚀

Engineering the parameter space

Emergent Methods

Defining the objective

- A clean and well defined parameter space:
  - enables **clustering** of news topics across **diverse perspectives**
  - represents **entities**, especially those originating from small demographics
  - “normalizes” for **language differences**

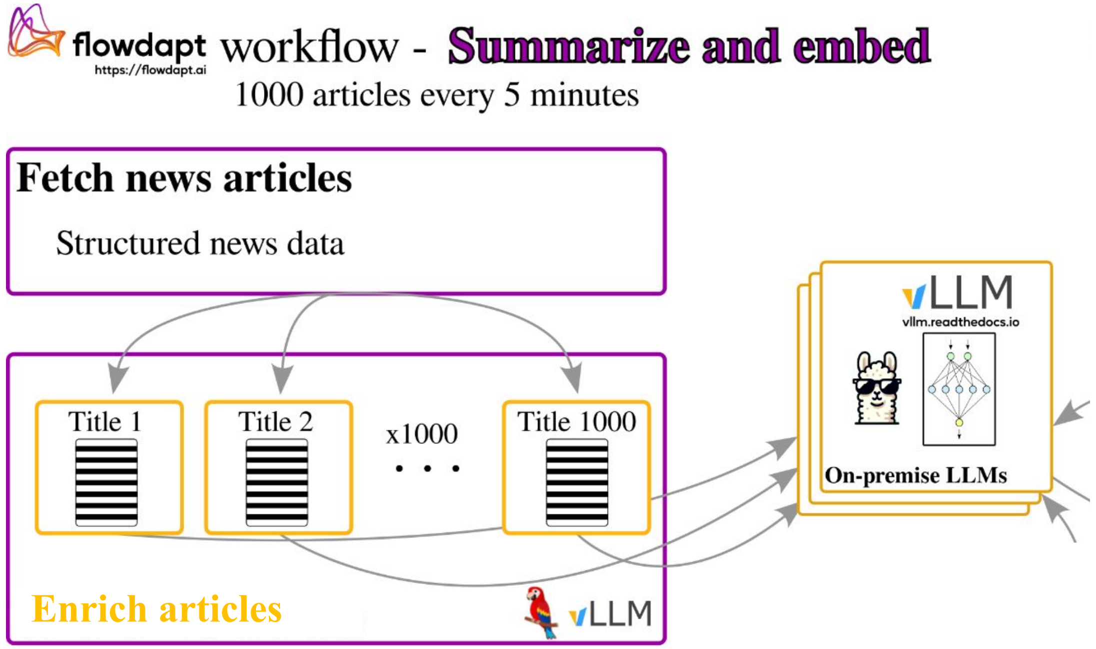


2024-05-07 Unstructured Data Meetup: Multilingual narrative tracking in the news - real-time experiments

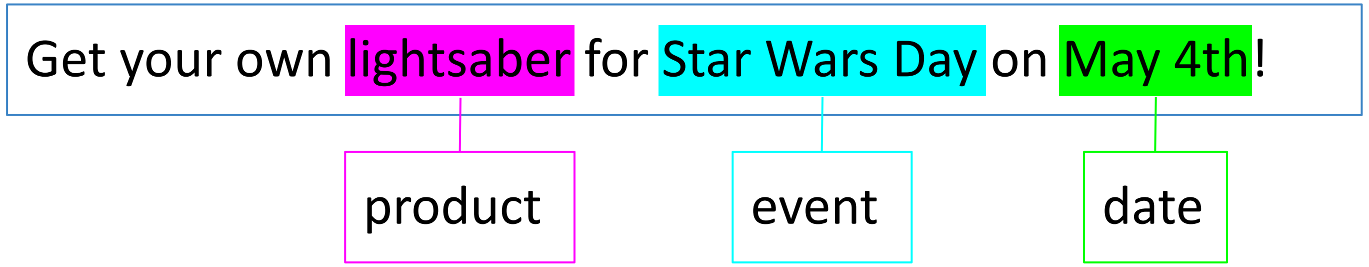
Preparing the embedding

- LLM **Enrich articles**
  - Translate
  - Summarize
  - Extract keywords, classification, and sentiment

Flexibility -> adaptability -> product opportunities


flowdapt https://flowdapt.ai workflow - **Summarize and embed**
1000 articles every 5 minutes
Fetch news articles
Structured news data
Title 1
Title 2
x1000
Title 1000
Enrich articles
vLLM
vLLM
vllm.readthedocs.io
On-premise LLMs

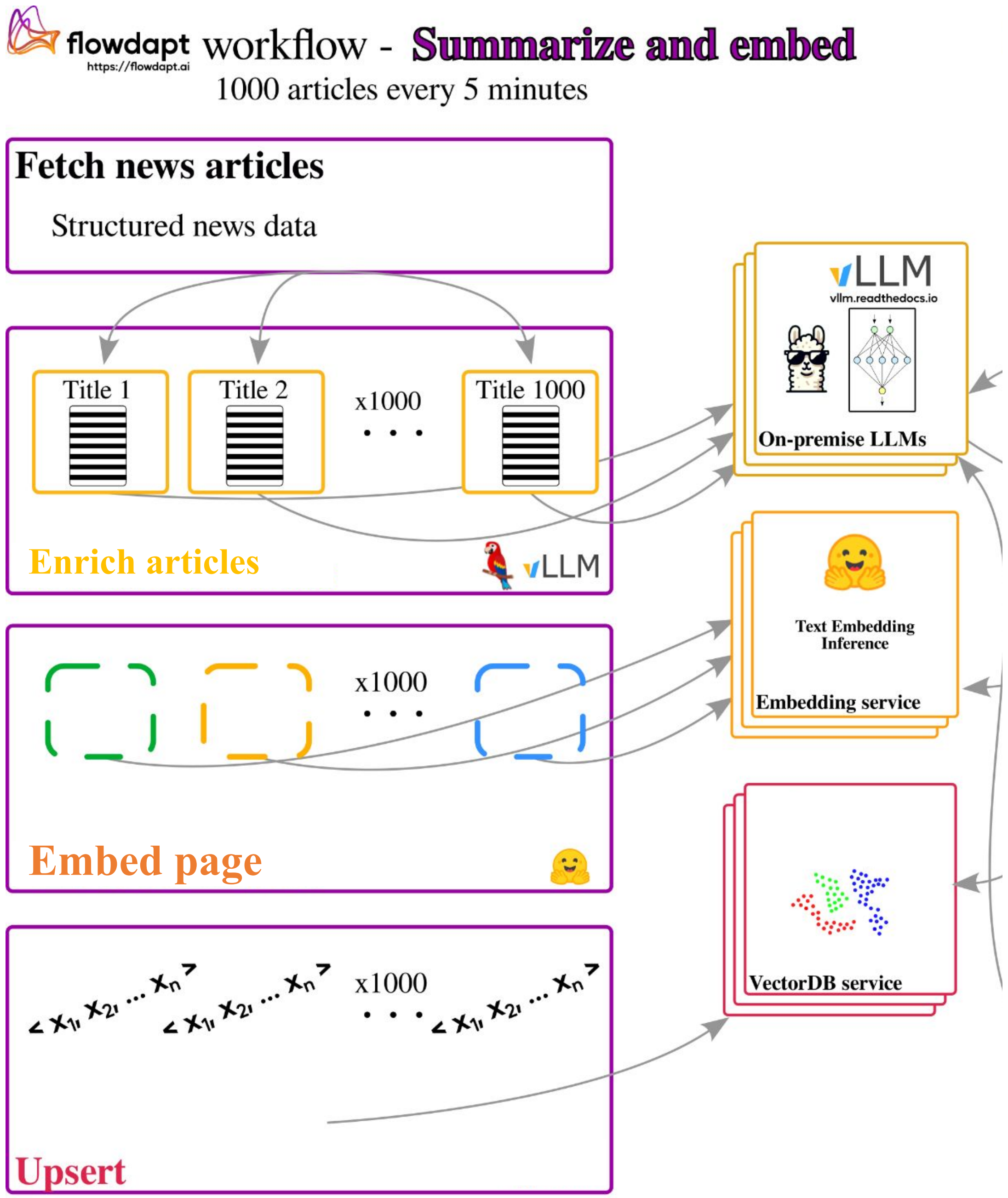
- Fine-tuned LM (GLiNER-news) **Enrich articles**
  - Generalist and lightweight entity extraction
  - Based model by Zaratiana et al. (2023)



Embedding and storing the page

- Build and embed page **Build/Embed page**
  - Model choice affects everything:
    - Retrieval speed/quality
    - Storage costs
    - Clustering compute costs
  - Doc structure vs expected Query structure

- VectorDB **Upsert**
  - The beating heart of the architecture:
    - Robustness
    - Parallelizability
    - Metadata filtering flexibility
    - Quantization
    - Performance



Tracking narratives in our parameter space

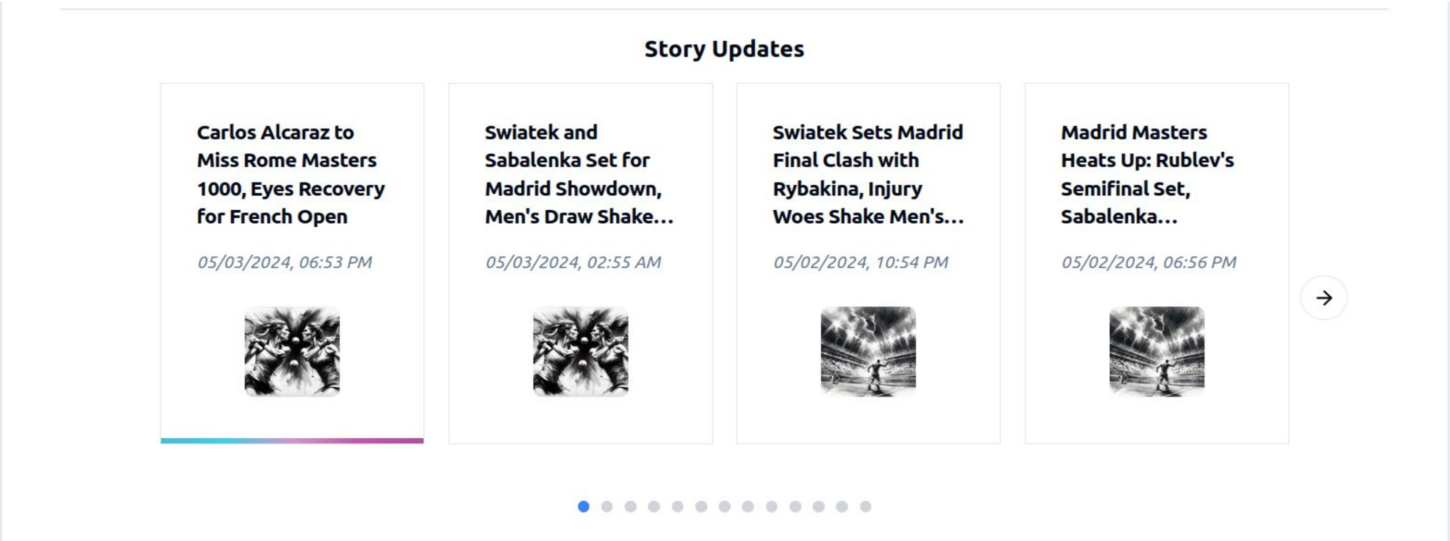
Emergent Methods

2024-05-07 Unstructured Data Meetup: Multilingual narrative tracking in the news - real-time experiments

Emergent Methods
www.emergentmethods.ai

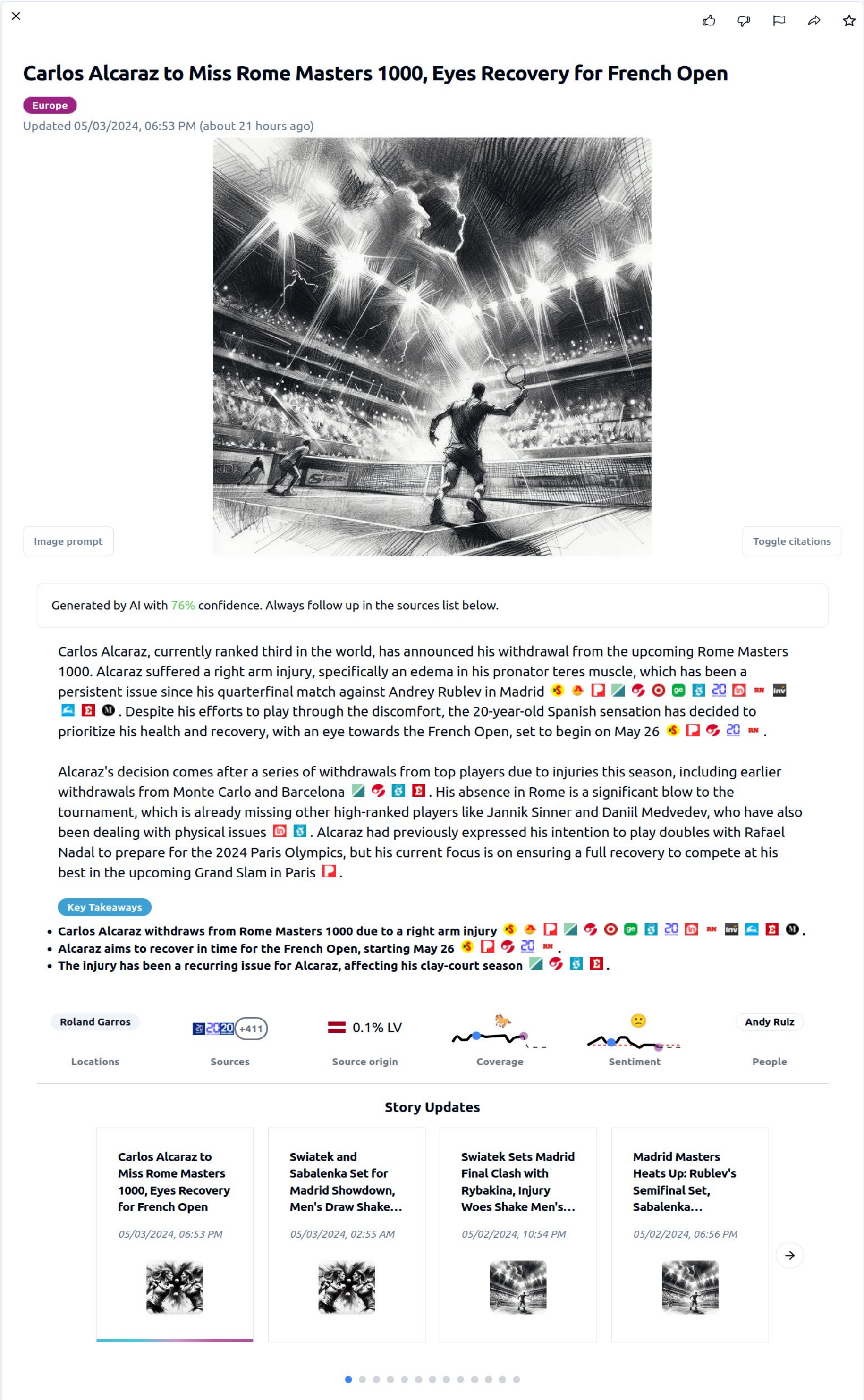
What is a news narrative?

- 🗞️ A series of related news reports
- 👬👫 Multiple points of view
- ❌ Errors - accidental and purposeful



Story Updates

Carlos Alcaraz to Miss Rome Masters 1000, Eyes Recovery for French Open	Swiatek and Sabalenka Set for Madrid Showdown, Men's Draw Shake...	Swiatek Sets Madrid Final Clash with Rybakina, Injury Woes Shake Men's...	Madrid Masters Heats Up: Rublev's Semifinal Set, Sabalenka...
05/03/2024, 06:53 PM	*05/03/2024, 02:55 AM*	*05/02/2024, 10:54 PM*	*05/02/2024, 06:56 PM*



Carlos Alcaraz to Miss Rome Masters 1000, Eyes Recovery for French Open

Updated 05/03/2024, 06:53 PM (about 21 hours ago)

Generated by AI with 76% confidence. Always follow up in the sources list below.

Carlos Alcaraz, currently ranked third in the world, has announced his withdrawal from the upcoming Rome Masters 1000. Alcaraz suffered a right arm injury, specifically an edema in his pronator teres muscle, which has been a persistent issue since his quarterfinal match against Andrey Rublev in Madrid . Despite his efforts to play through the discomfort, the 20-year-old Spanish sensation has decided to prioritize his health and recovery, with an eye towards the French Open, set to begin on May 26 .

Alcaraz's decision comes after a series of withdrawals from top players due to injuries this season, including earlier withdrawals from Monte Carlo and Barcelona . His absence in Rome is a significant blow to the tournament, which is already missing other high-ranked players like Jannik Sinner and Daniil Medvedev, who have also been dealing with physical issues . Alcaraz had previously expressed his intention to play doubles with Rafael Nadal to prepare for the 2024 Paris Olympics, but his current focus is on ensuring a full recovery to compete at his best in the upcoming Grand Slam in Paris .

Key Takeaways

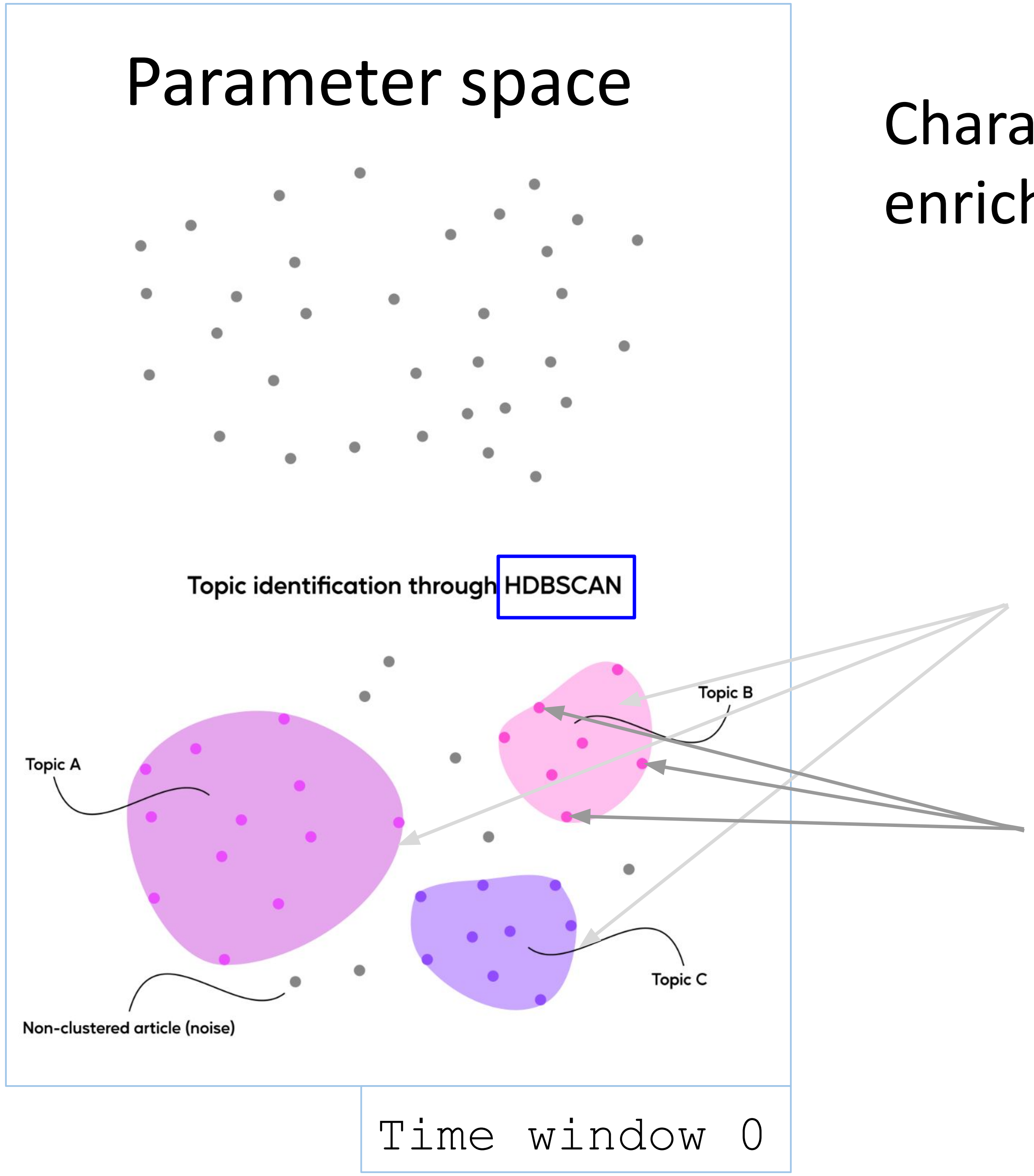
- **Carlos Alcaraz withdraws from Rome Masters 1000 due to a right arm injury .**
- **Alcaraz aims to recover in time for the French Open, starting May 26 .**
- **The injury has been a recurring issue for Alcaraz, affecting his clay-court season .**

Roland Garros	+411	0.1% LV			Andy Ruiz
Locations	Sources	Source origin	Coverage	Sentiment	People

Story Updates

Carlos Alcaraz to Miss Rome Masters 1000, Eyes Recovery for French Open	Swiatek and Sabalenka Set for Madrid Showdown, Men's Draw Shake...	Swiatek Sets Madrid Final Clash with Rybakina, Injury Woes Shake Men's...	Madrid Masters Heats Up: Rublev's Semifinal Set, Sabalenka...
05/03/2024, 06:53 PM	*05/03/2024, 02:55 AM*	*05/02/2024, 10:54 PM*	*05/02/2024, 06:56 PM*

Finding related news reports


Parameter space
Topic identification through HDBSCAN
Topic A
Topic B
Topic C
Non-clustered article (noise)
Time window 0

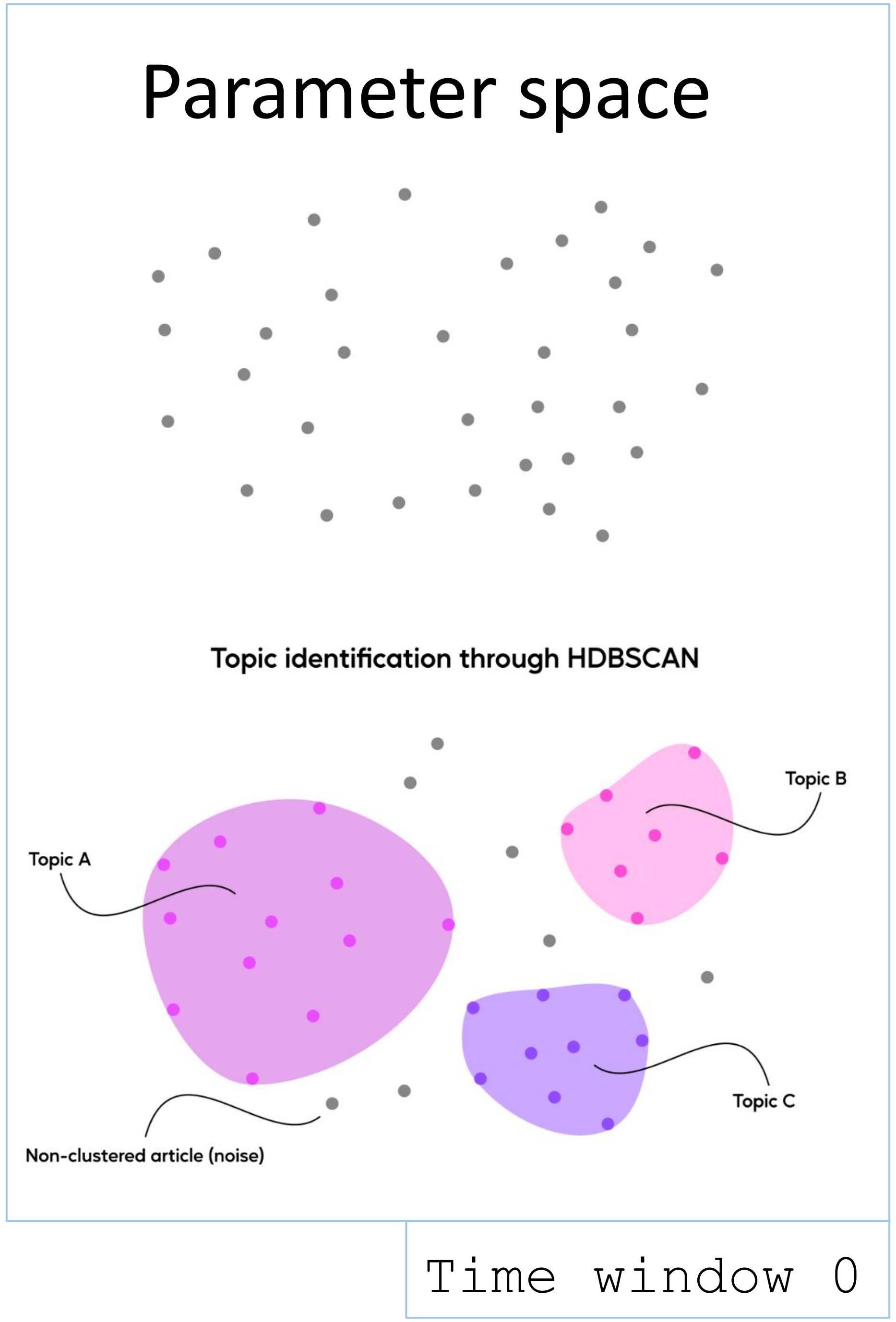
Characterized by our embedded enrichments

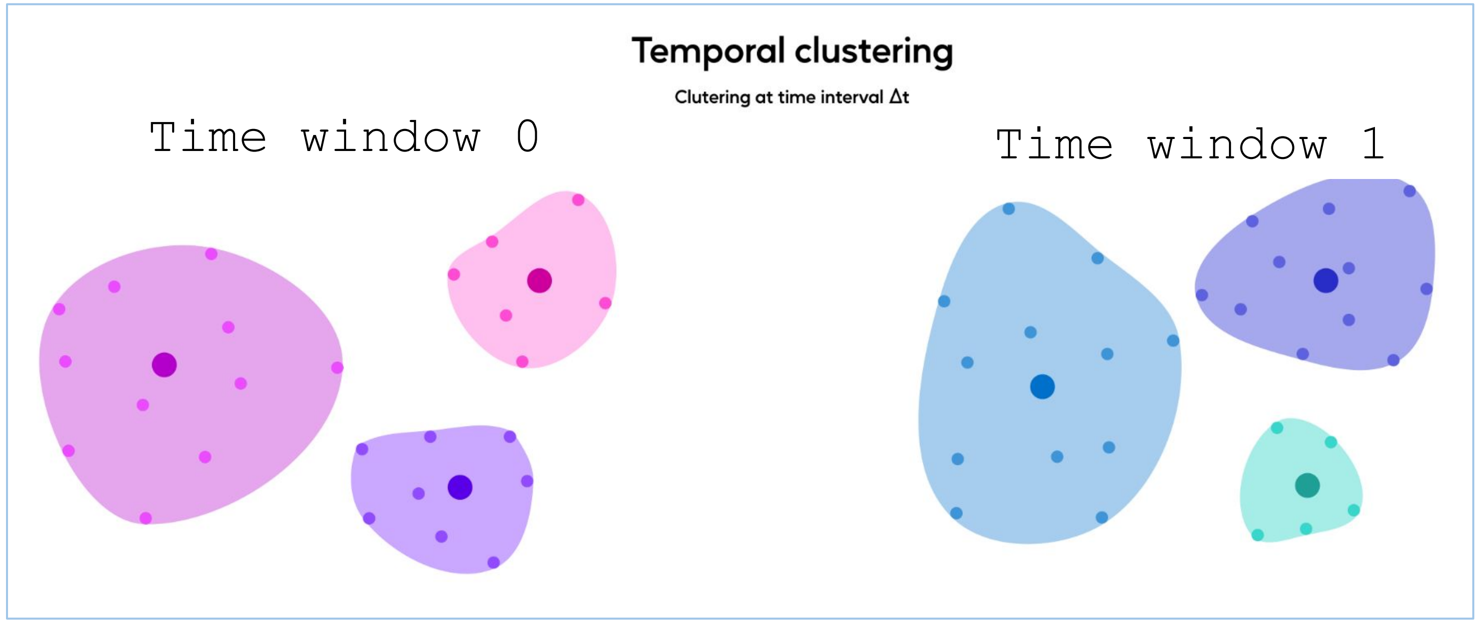
Clusters of semantic similarity

Multiple countries
Multiple languages
Multiple sources
Competing perspectives

2024-05-07 Unstructured Data Meetup: Multilingual narrative tracking in the news - real-time experiments

Identifying the series of events


Parameter space
Topic identification through HDBSCAN
Topic A
Topic B
Topic C
Non-clustered article (noise)
Time window 0

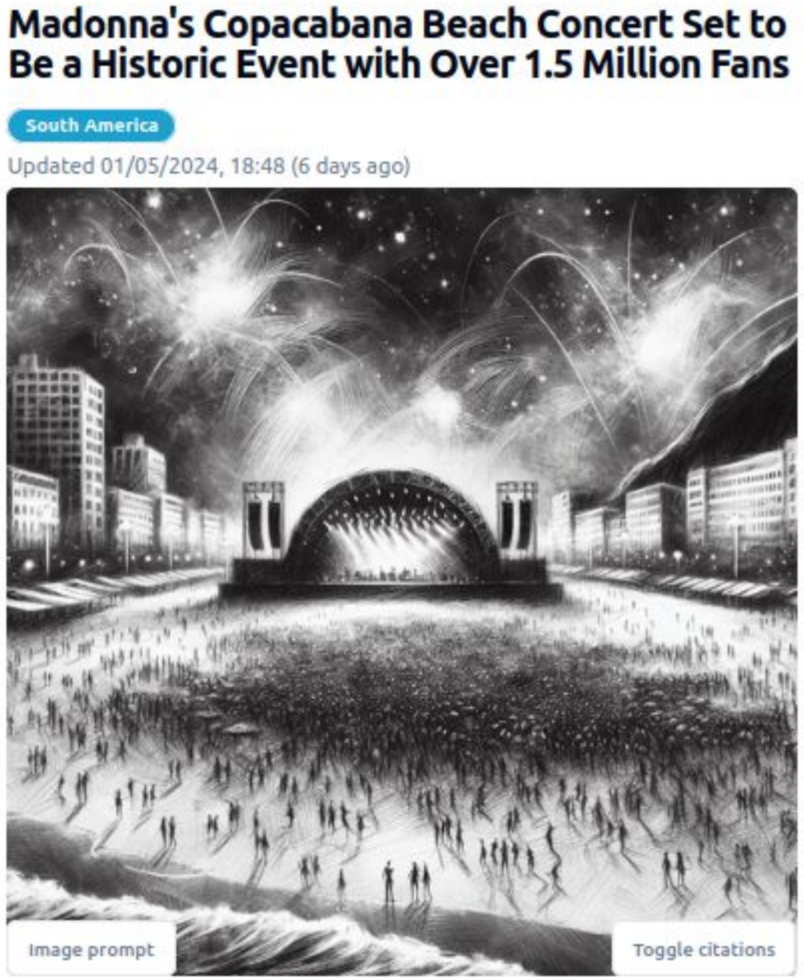


Cluster connection methods

- Adaptively (re)train one binary classifier per cluster
- Track medoid/centroid drift with hyper spheres
- Use overlap clustering techniques

Long range niche tracking

May 1, 2024

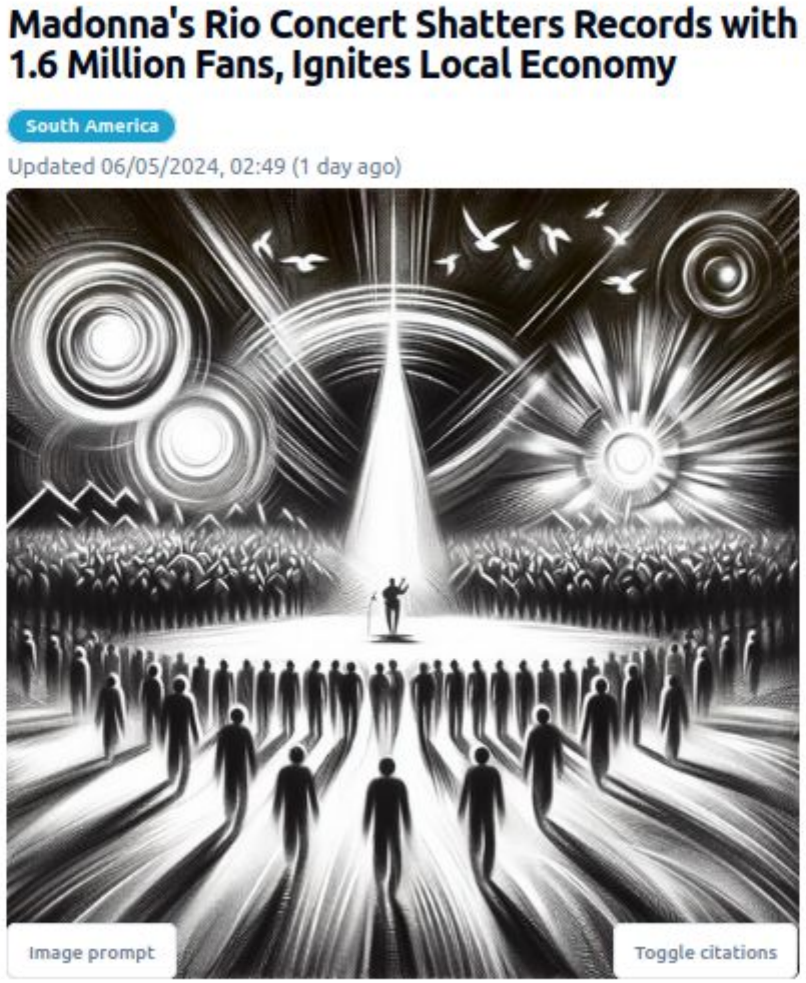


Madonna's Copacabana Beach Concert Set to Be a Historic Event with Over 1.5 Million Fans

South America

Updated 01/05/2024, 18:48 (6 days ago)

May 6, 2024



Madonna's Rio Concert Shatters Records with 1.6 Million Fans, Ignites Local Economy

South America

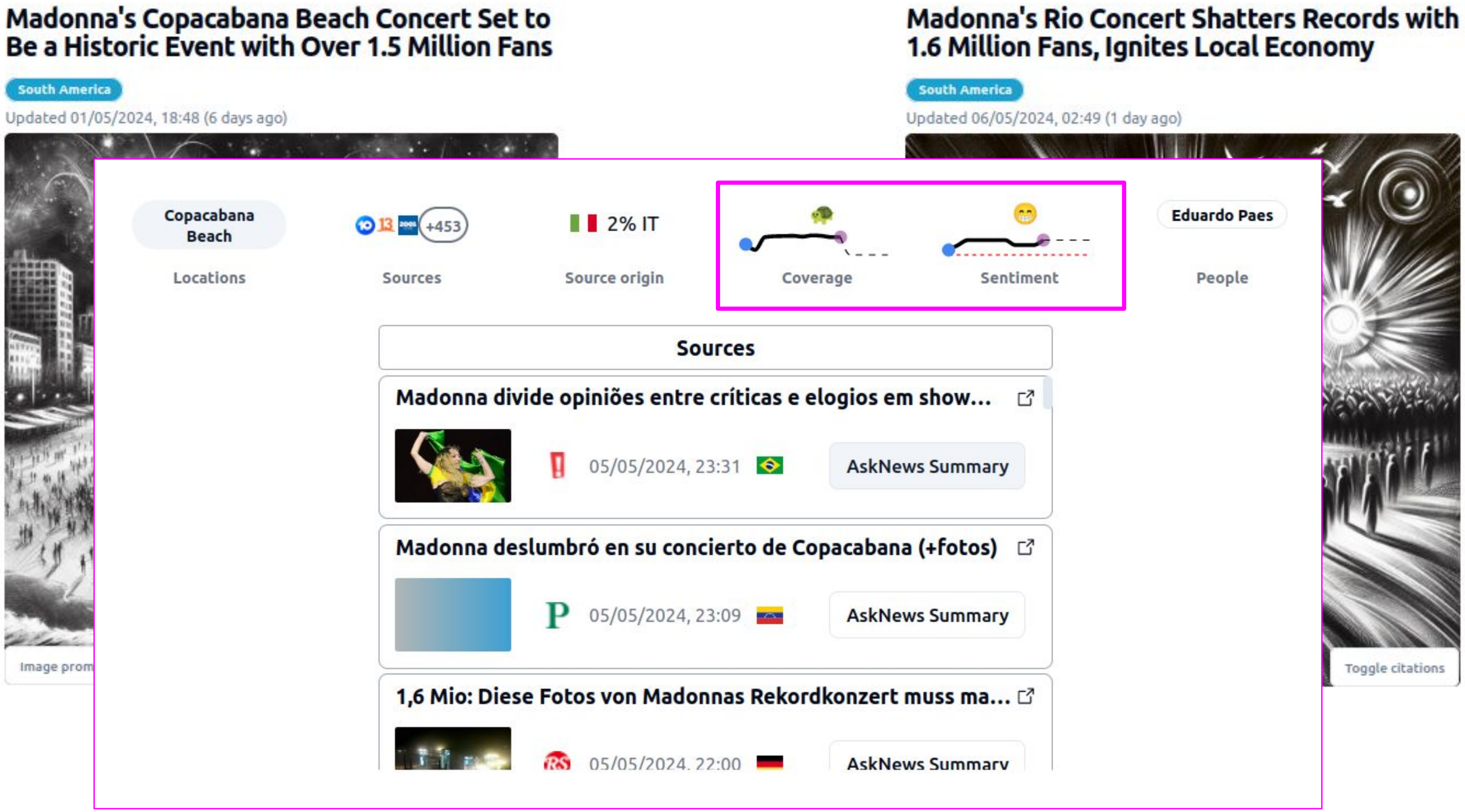
Updated 06/05/2024, 02:49 (1 day ago)

2024-05-07 Unstructured Data Meetup: Multilingual narrative tracking in the news - real-time experiments

Emergent Methods
www.emergentmethods.ai

Long range niche tracking

May 1, 2024 → May 6, 2024



Madonna's Copacabana Beach Concert Set to Be a Historic Event with Over 1.5 Million Fans

South America

Updated 01/05/2024, 18:48 (6 days ago)

Image prom

Madonna's Rio Concert Shatters Records with 1.6 Million Fans, Ignites Local Economy

South America

Updated 06/05/2024, 02:49 (1 day ago)

Toggle citations

Copacabana Beach	+453	2% IT			Eduardo Paes
Locations	Sources	Source origin	Coverage	Sentiment	People

Sources

Madonna divide opiniões entre críticas e elogios em show...
05/05/2024, 23:31 — AskNews Summary

Madonna deslumbró en su concierto de Copacabana (+fotos)
05/05/2024, 23:09 — AskNews Summary

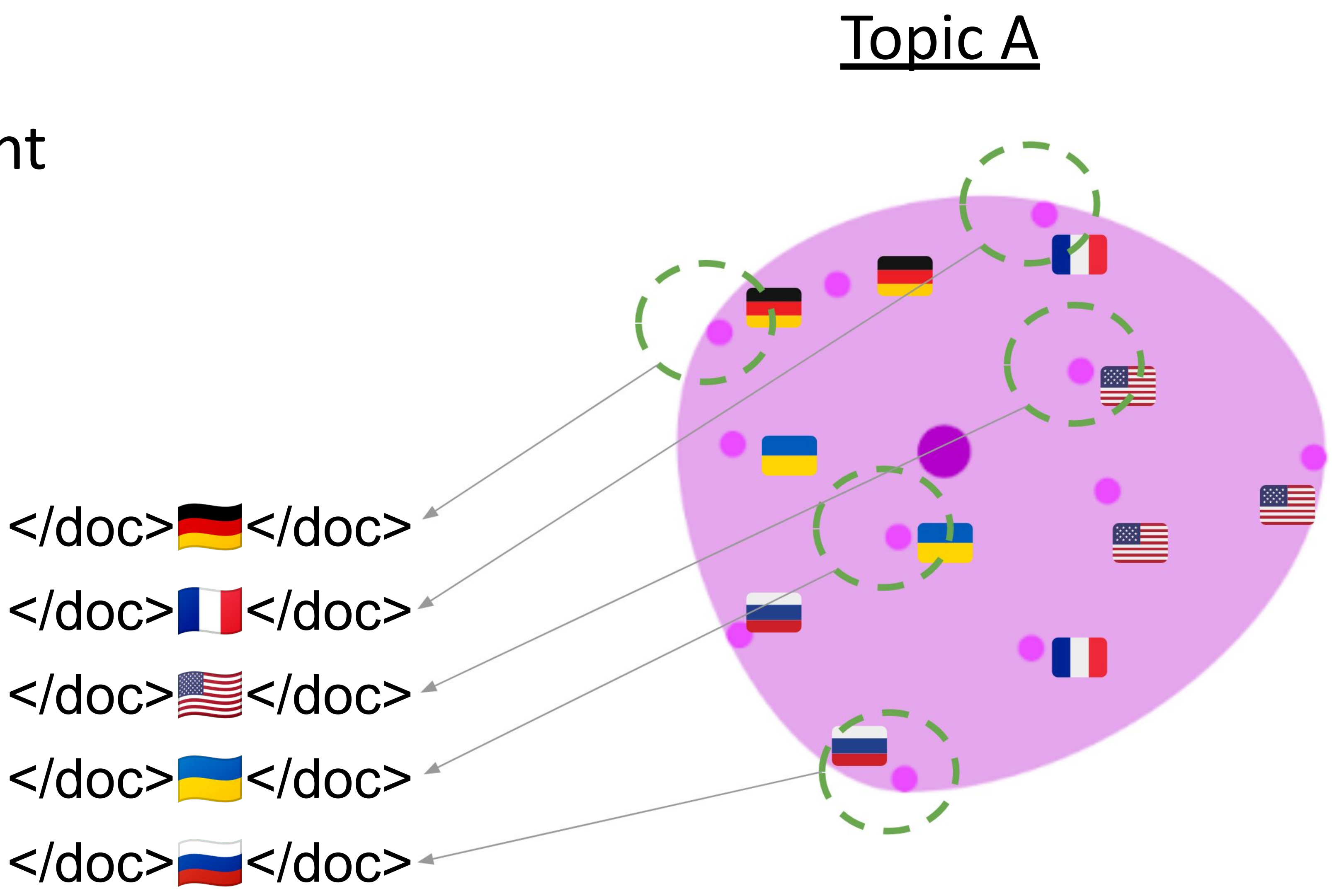
1,6 Mio: Diese Fotos von Madonnas Rekordkonzert muss ma...
05/05/2024, 22:00 — AskNews Summary

Context polishing (enforcing diversity)

- Pruning a single cluster
  - diversity enforcement
- Confirming continuity
- Reranking the cluster

Prompt engineering

- Document formatting
- Citation control
- Journalistic guidelines



Topic A

</doc>🇩🇪</doc>

</doc>🇫🇷</doc>

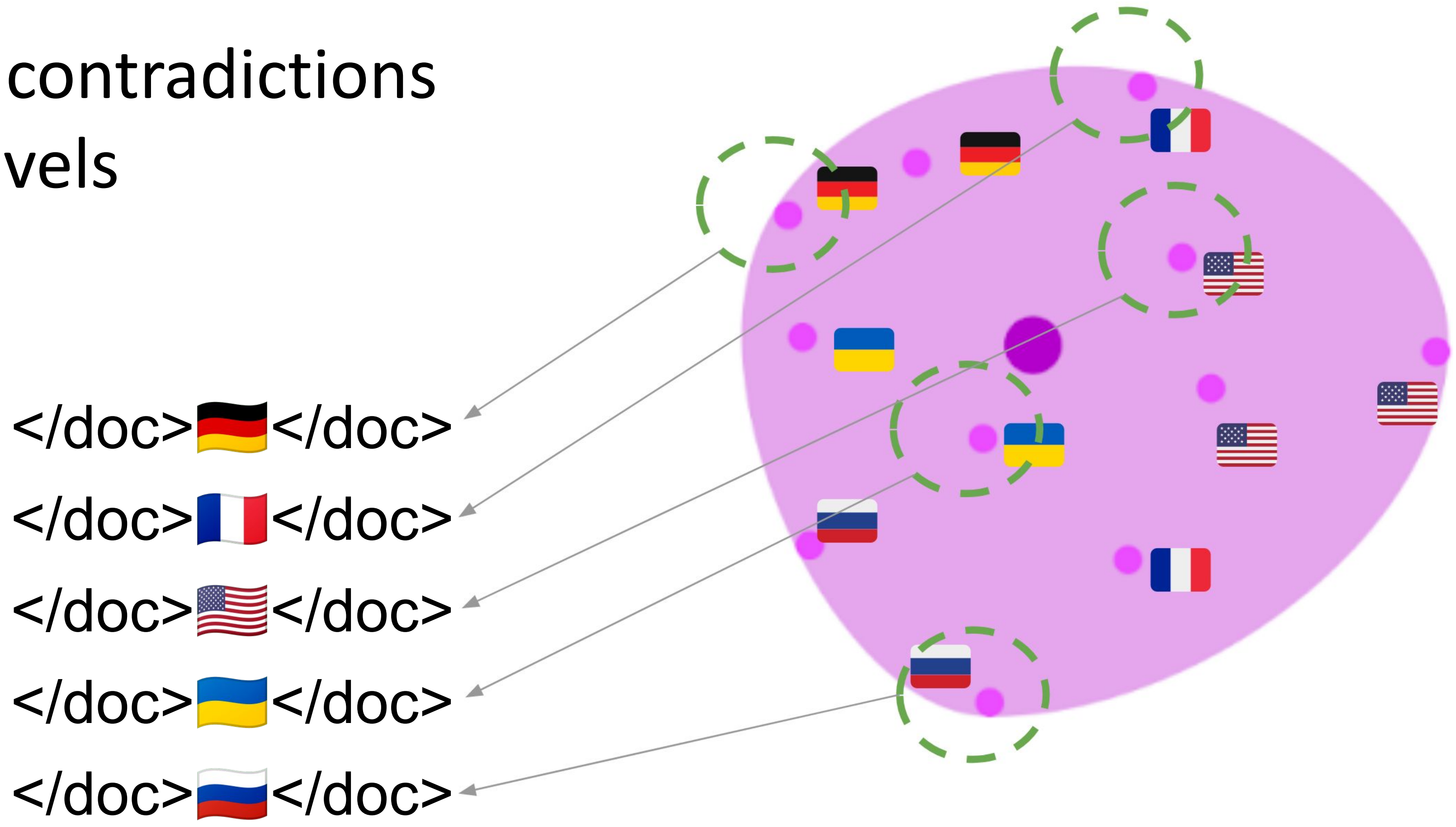
</doc>🇺🇸</doc>

</doc>🇺🇦</doc>

</doc>🇷🇺</doc>

Report cluster alignment (quantifying diversity)

- Identify alignment and contradictions
- Compute confidence levels


Topic A

</doc>🇩🇪</doc>

</doc>🇫🇷</doc>

</doc>🇺🇸</doc>

</doc>🇺🇦</doc>

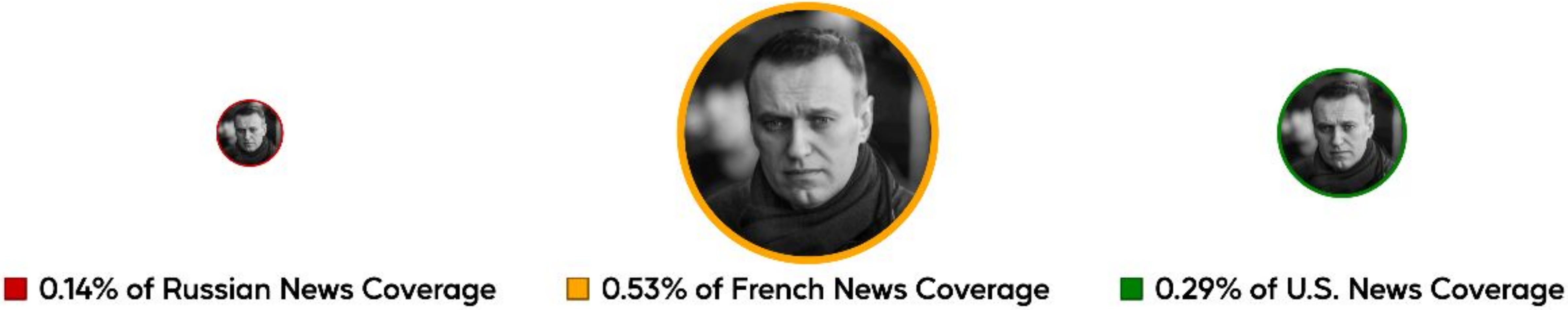
</doc>🇷🇺</doc>

Tracking the death of Alexei Navalny

Emergent Methods

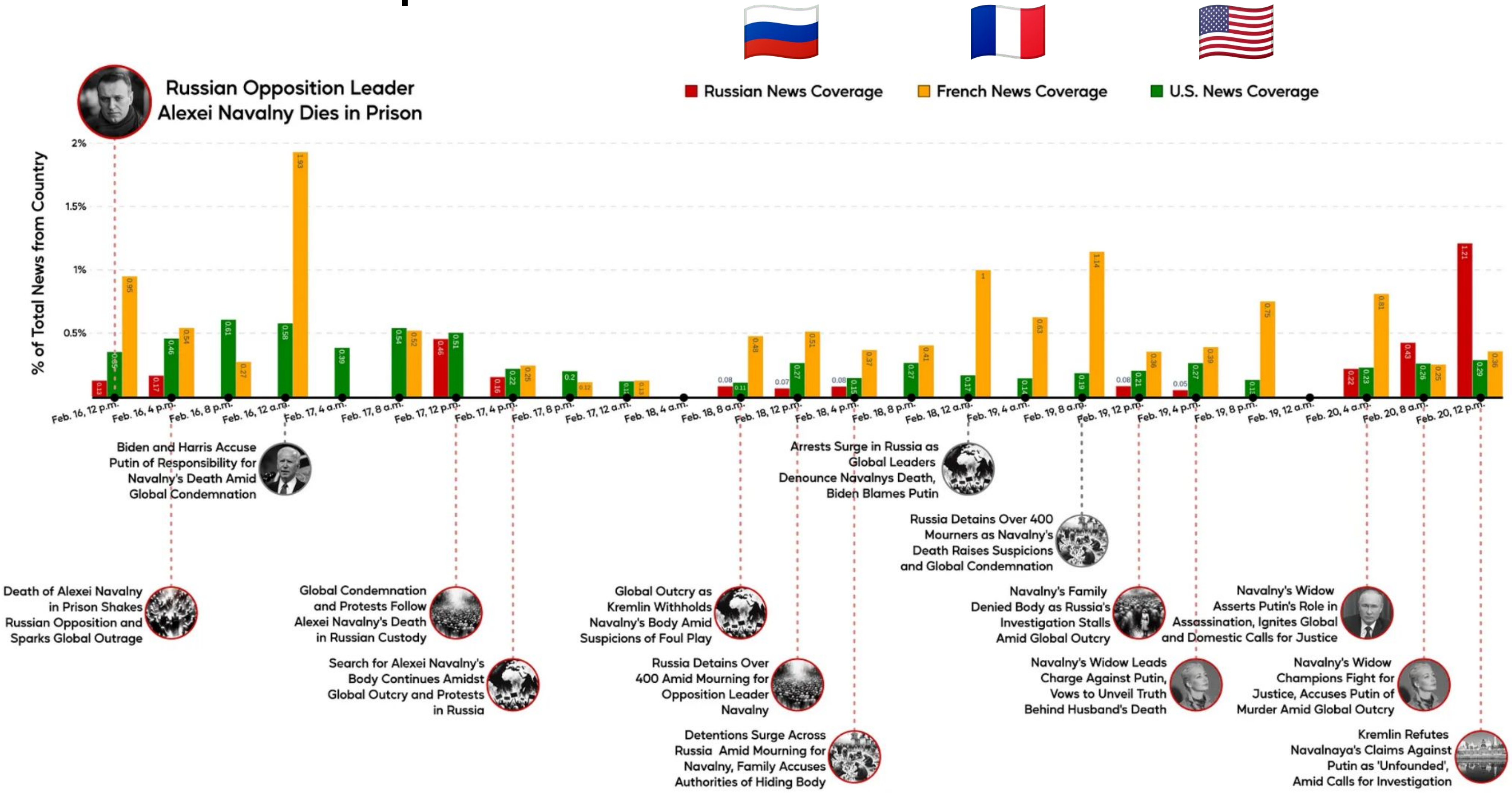
🇺🇸 🇫🇷 🇷🇺 comparison

Percentage of total country news coverage devoted to Navalny narrative



🇷🇺 🇫🇷 🇺🇸

🇺🇸 🇫🇷 🇷🇺 comparison



Biden and Harris Accuse Putin of Responsibility for Navalny's Death Amid Global Condemnation

Death of Alexei Navalny in Prison Shakes Russian Opposition and Sparks Global Outrage

Global Condemnation and Protests Follow Alexei Navalny's Death in Russian Custody

Search for Alexei Navalny's Body Continues Amidst Global Outcry and Protests in Russia

Global Outcry as Kremlin Withholds Navalny's Body Amid Suspicions of Foul Play

Russia Detains Over 400 Amid Mourning for Opposition Leader Navalny

Detentions Surge Across Russia Amid Mourning for Navalny, Family Accuses Authorities of Hiding Body

Arrests Surge in Russia as Global Leaders Denounce Navalnys Death, Biden Blames Putin

Russia Detains Over 400 Mourners as Navalny's Death Raises Suspicions and Global Condemnation

Navalny's Family Denied Body as Russia's Investigation Stalls Amid Global Outcry

Navalny's Widow Leads Charge Against Putin, Vows to Unveil Truth Behind Husband's Death

Navalny's Widow Asserts Putin's Role in Assassination, Ignites Global and Domestic Calls for Justice

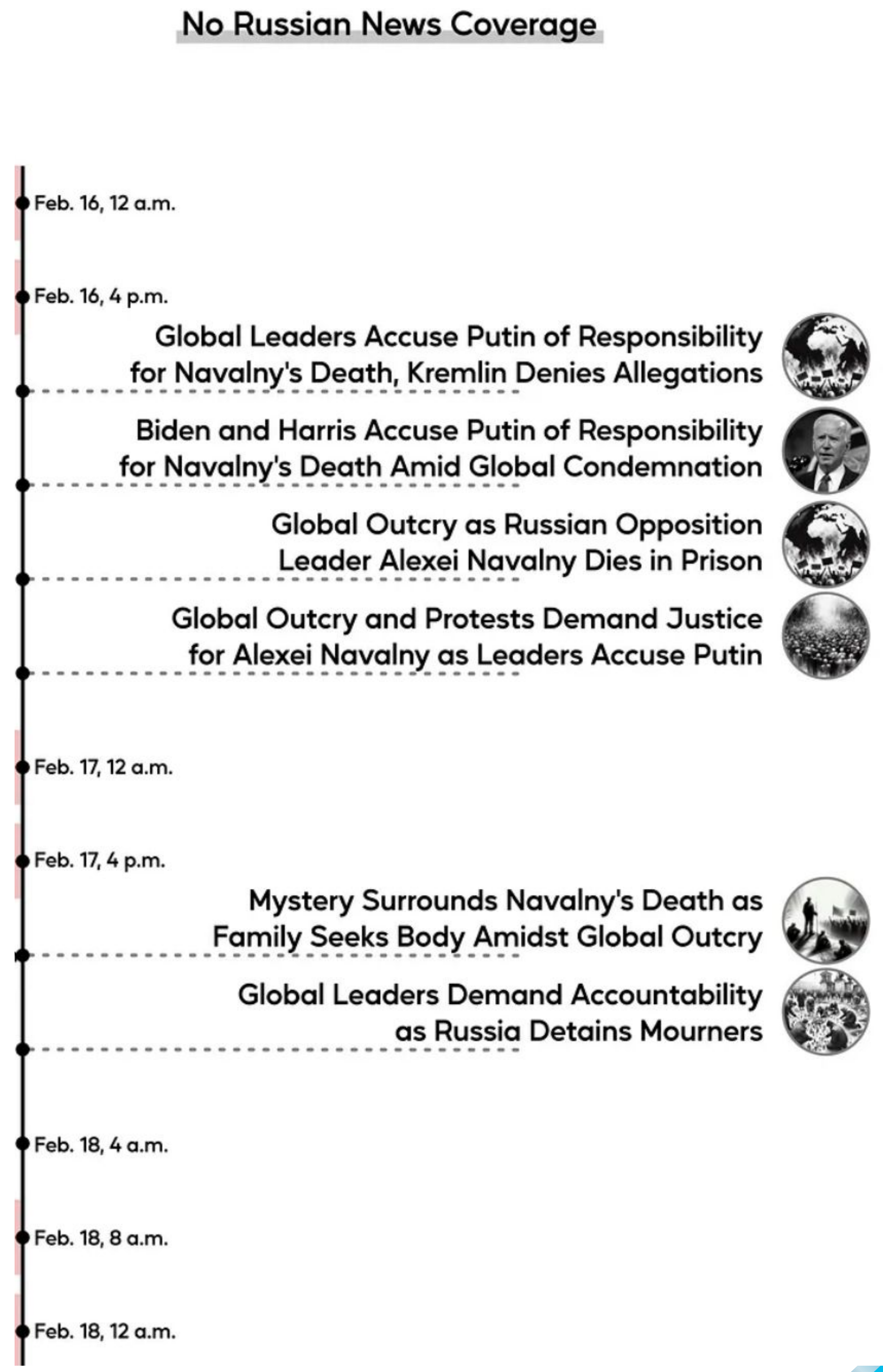
Navalny's Widow Champions Fight for Justice, Accuses Putin of Murder Amid Global Outcry

Kremlin Refutes Navalnaya's Claims Against Putin as 'Unfounded', Amid Calls for Investigation

Russian, French, and U.S. news coverage of important events following the death of Alexei Navalny. Timestamps are in UTC.

Uncovering non-reporting

- Which aspects of the narrative were least reported by Russian sources?



No Russian News Coverage

Feb. 16, 12 a.m.

Feb. 16, 4 p.m.

Global Leaders Accuse Putin of Responsibility for Navalny's Death, Kremlin Denies Allegations

Biden and Harris Accuse Putin of Responsibility for Navalny's Death Amid Global Condemnation

Global Outcry as Russian Opposition Leader Alexei Navalny Dies in Prison

Global Outcry and Protests Demand Justice for Alexei Navalny as Leaders Accuse Putin

Feb. 17, 12 a.m.

Feb. 17, 4 p.m.

Mystery Surrounds Navalny's Death as Family Seeks Body Amidst Global Outcry

Global Leaders Demand Accountability as Russia Detains Mourners

Feb. 18, 4 a.m.

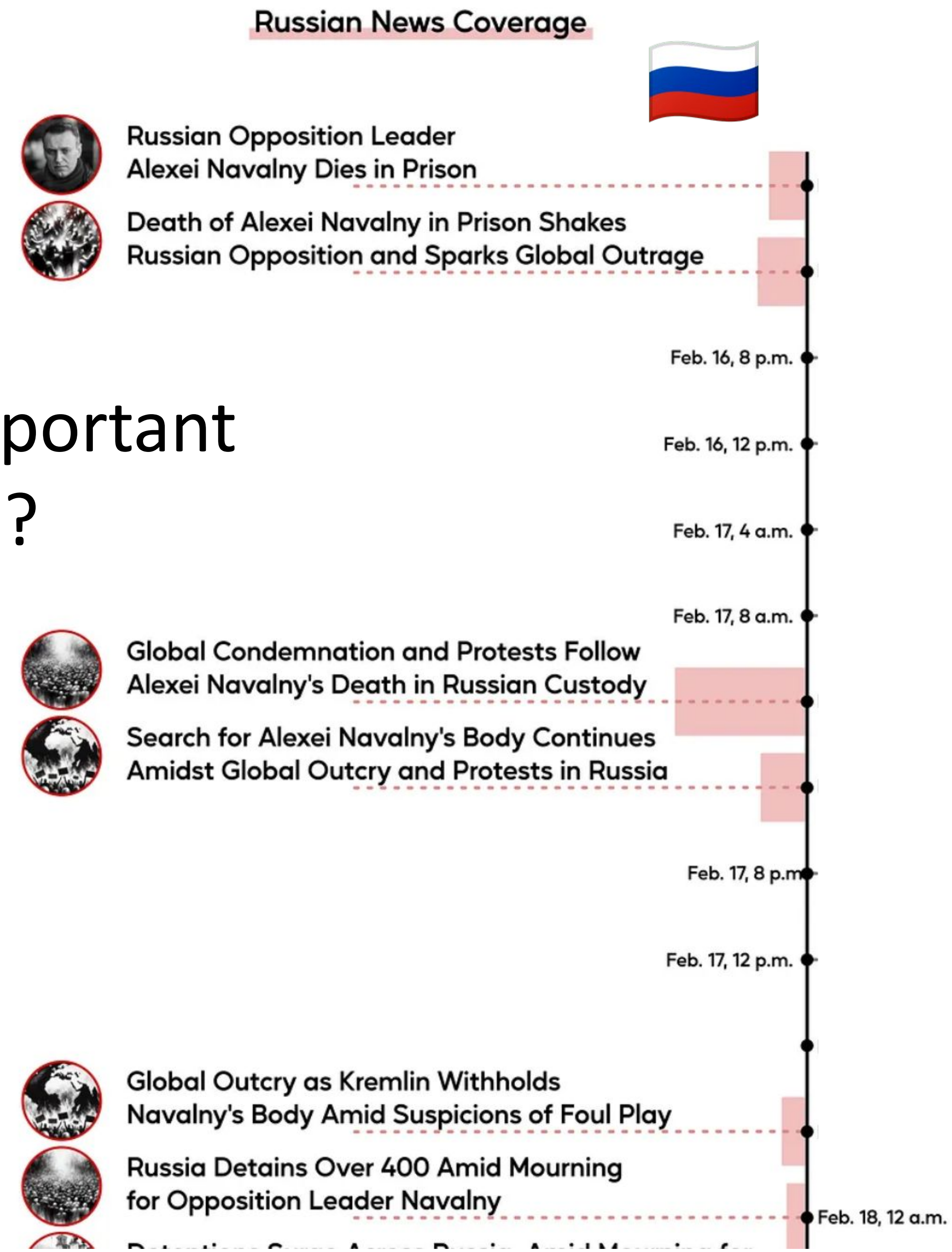
Feb. 18, 8 a.m.

Feb. 18, 12 a.m.

Emergent Methods
www.emergentmethods.ai

Uncovering non-reporting

- Which details were important to the Russian sources?



Russian News Coverage

Russian Opposition Leader Alexei Navalny Dies in Prison

Death of Alexei Navalny in Prison Shakes Russian Opposition and Sparks Global Outrage

Global Condemnation and Protests Follow Alexei Navalny's Death in Russian Custody

Search for Alexei Navalny's Body Continues Amidst Global Outcry and Protests in Russia

Global Outcry as Kremlin Withholds Navalny's Body Amid Suspicions of Foul Play

Russia Detains Over 400 Amid Mourning for Opposition Leader Navalny

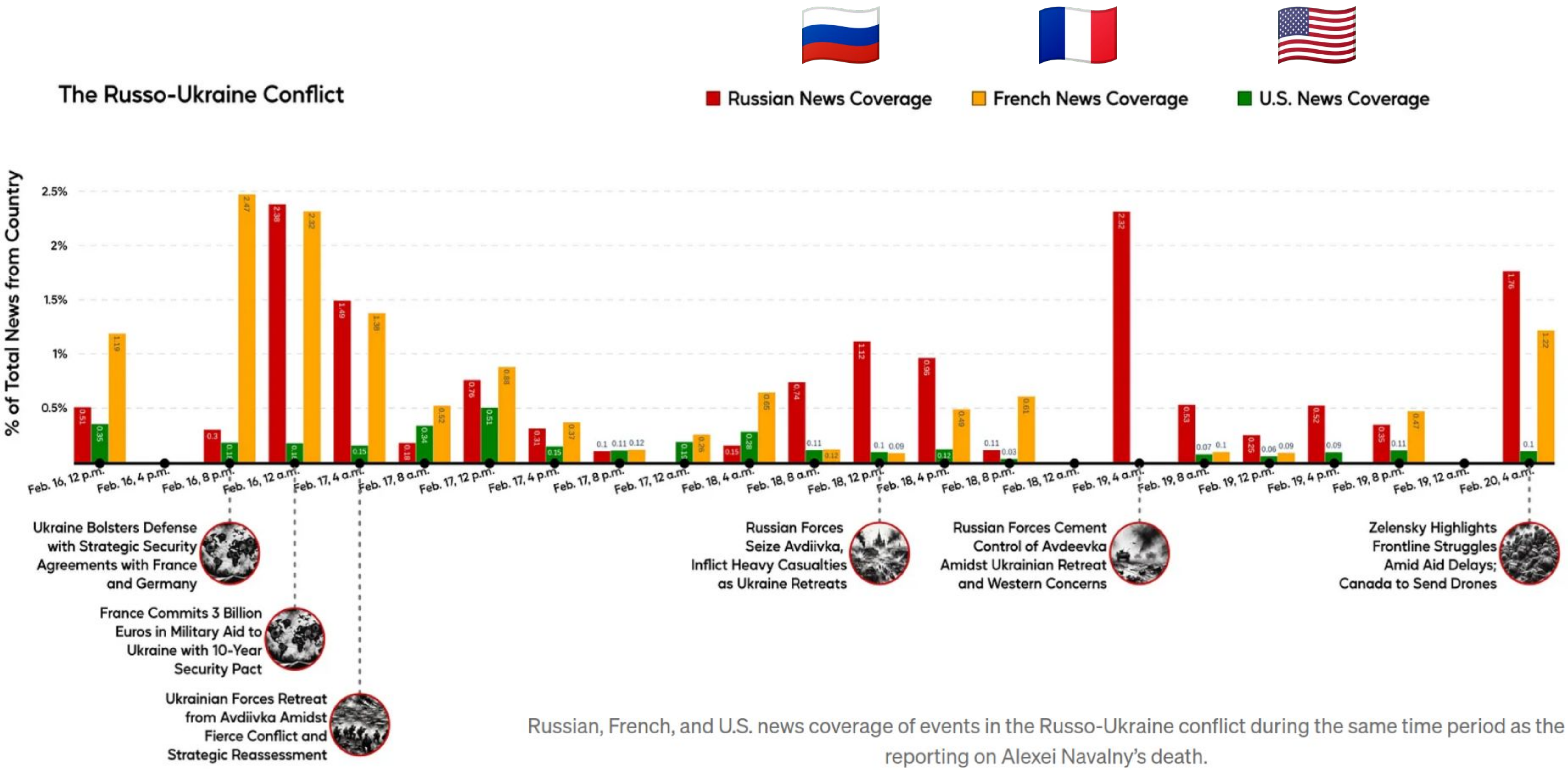
Detentions Surge Across Russia Amid Mourning for

2024-05-07 Unstructured Data Meetup: Multilingual narrative tracking in the news - real-time experiments

2024-05-07 Unstructured Data Meetup: Multilingual narrative tracking in the news - real-time experiments

Emergent Methods
www.emergentmethods.ai

Coverage of the Russo-Ukraine conflict



Ukraine Bolsters Defense with Strategic Security Agreements with France and Germany

France Commits 3 Billion Euros in Military Aid to Ukraine with 10-Year Security Pact

Ukrainian Forces Retreat from Avdiivka Amidst Fierce Conflict and Strategic Reassessment

Russian Forces Seize Avdiivka, Inflict Heavy Casualties as Ukraine Retreats

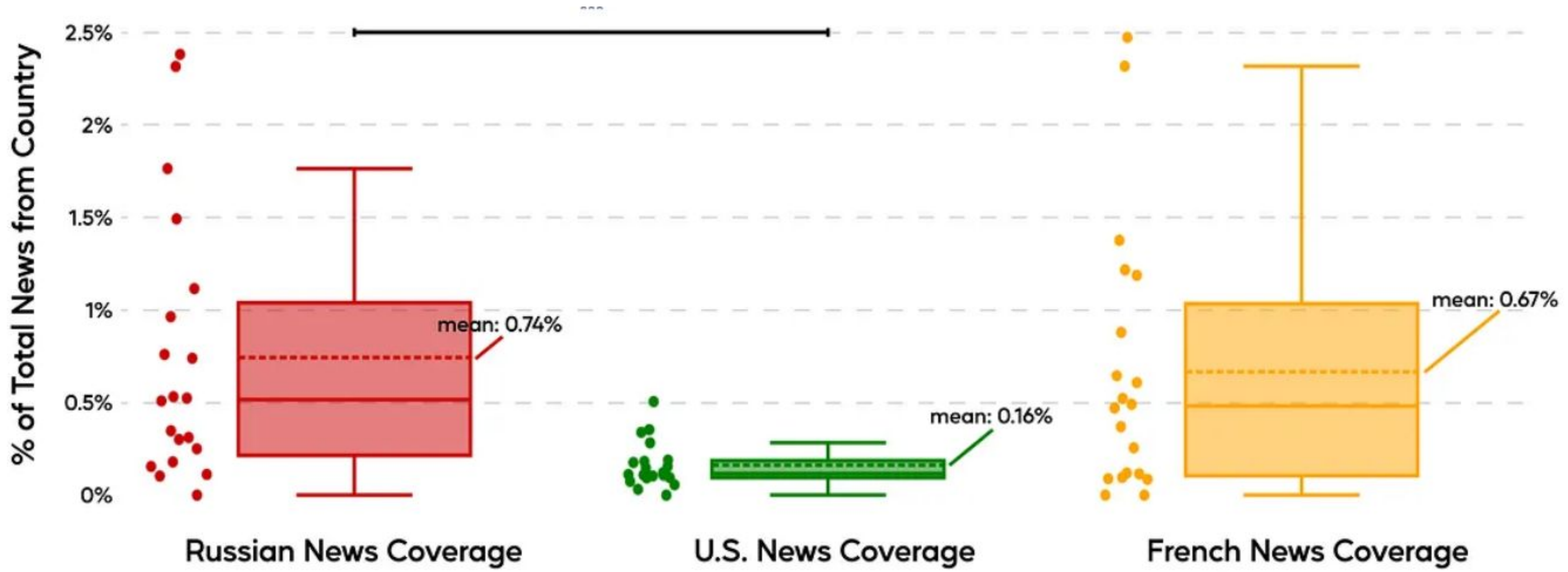
Russian Forces Cement Control of Avdeevka Amidst Ukrainian Retreat and Western Concerns

Zelensky Highlights Frontline Struggles Amid Aid Delays; Canada to Send Drones

Russian, French, and U.S. news coverage of events in the Russo-Ukraine conflict during the same time period as the reporting on Alexei Navalny's death.

Emergent Methods
www.emergentmethods.ai

Coverage of the Russo-Ukraine conflict

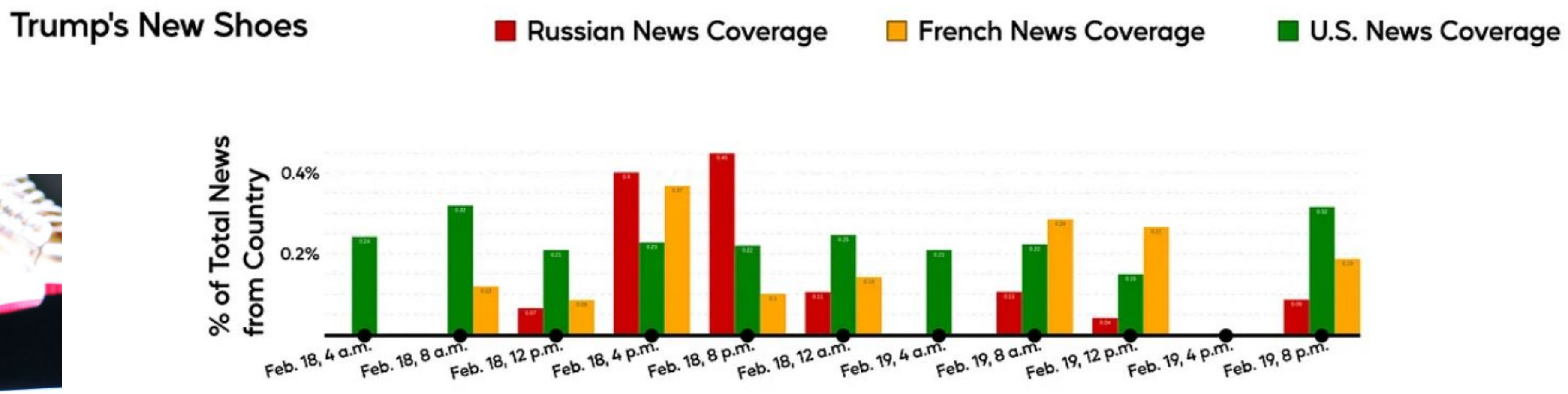


The Russian coverage of of the Russo-Ukraine conflict is the same as from French news outlets, while U.S. news coverage is significantly below Russian.

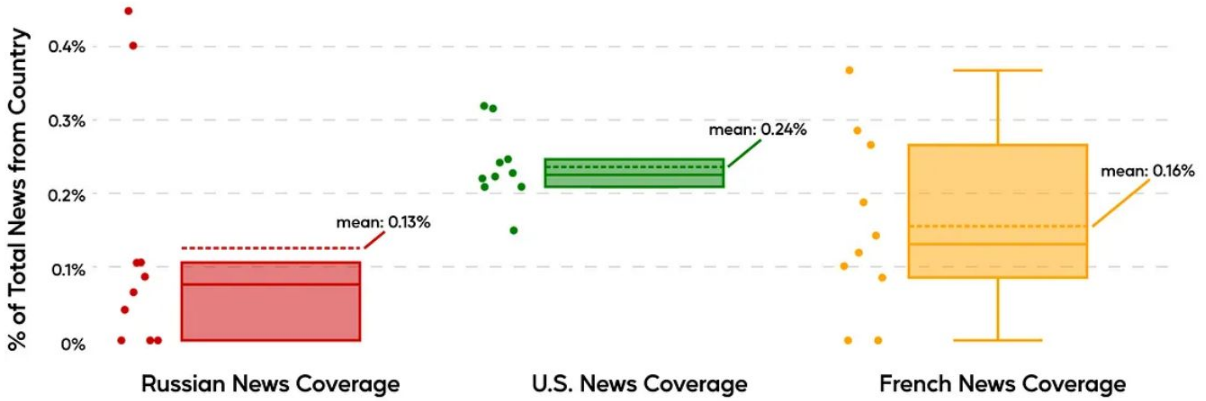
🇷🇺 🇺🇸 🇫🇷

2024-05-07 Unstructured Data Meetup: Multilingual narrative tracking in the news - real-time experiments

Coverage of Trump’s new shoes



News coverage of U.S. ex-president Donald Trump’s release of the “Never Surrender High-Tops” shoes.



Russian coverage of the release of U.S. ex-president Donald Trump’s new shoes matches that of France and U.S.

🇷🇺 🇺🇸 🇫🇷

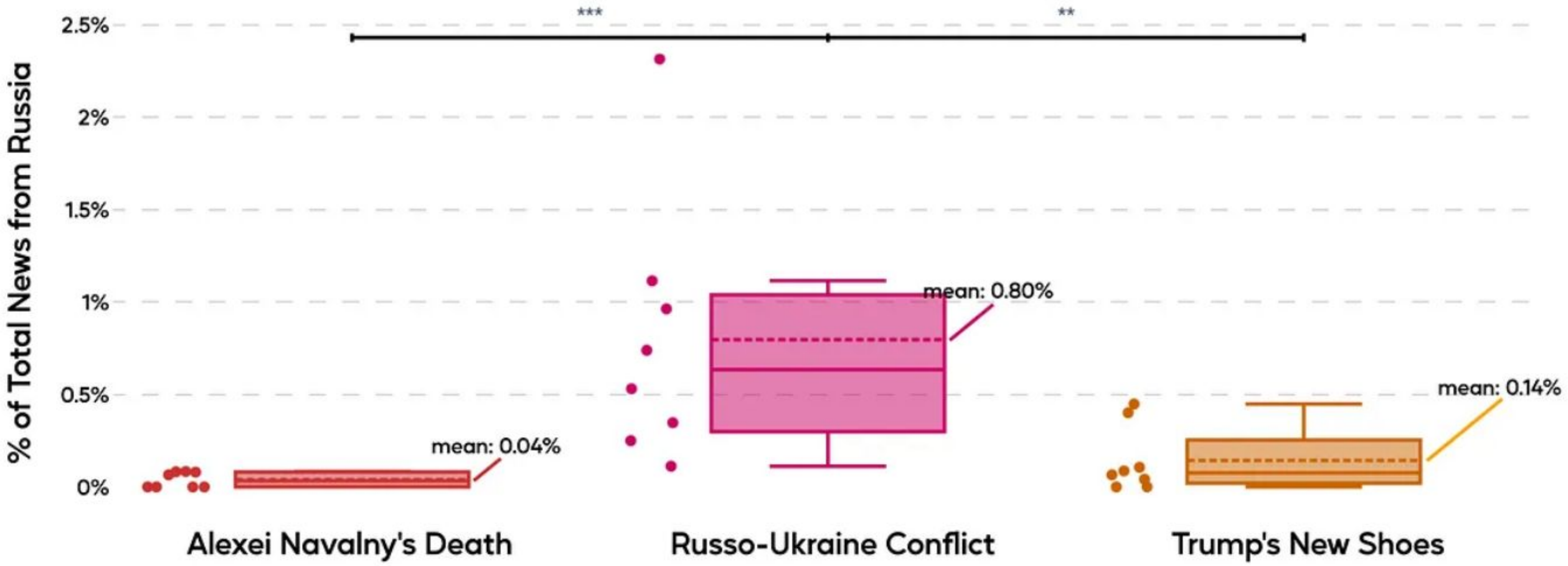
2024-05-07 Unstructured Data Meetup: Multilingual narrative tracking in the news - real-time experiments

Emergent Methods
www.emergentmethods.ai

Compared to other topics

Equivalent volume of Russian news coverage for:

- Trump’s new shoes
- Death of Navalvy’s death

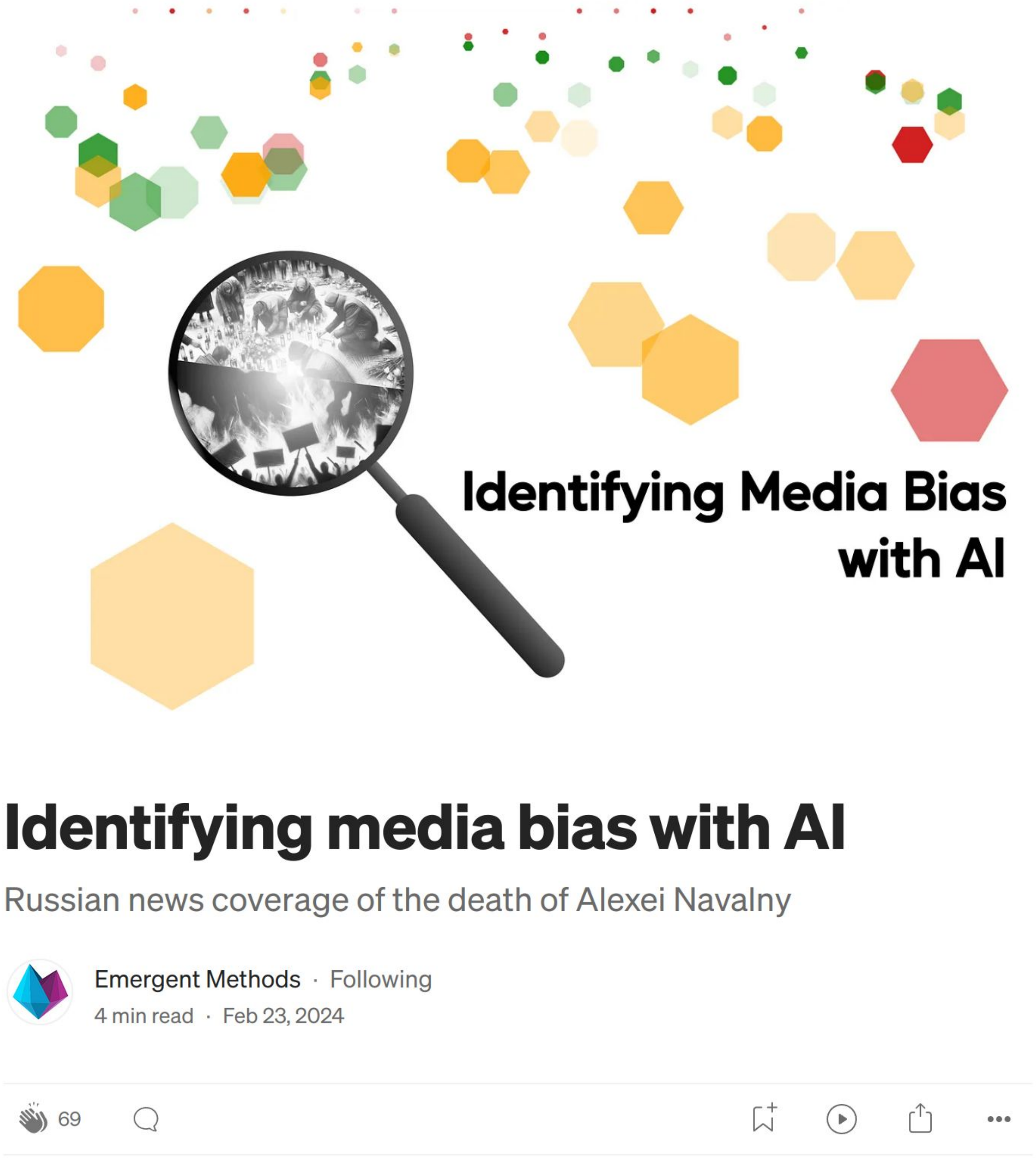


Russian news coverage of the death of Alexei Navalny, the Russo-Ukraine conflict, and Trumps new shoes during the topics’ overlapping time period.

Data and analysis available in [Google Colab]

Emergent Methods
www.emergentmethods.ai

Blog article available for more details



Identifying Media Bias with AI

Identifying media bias with AI

Russian news coverage of the death of Alexei Navalny

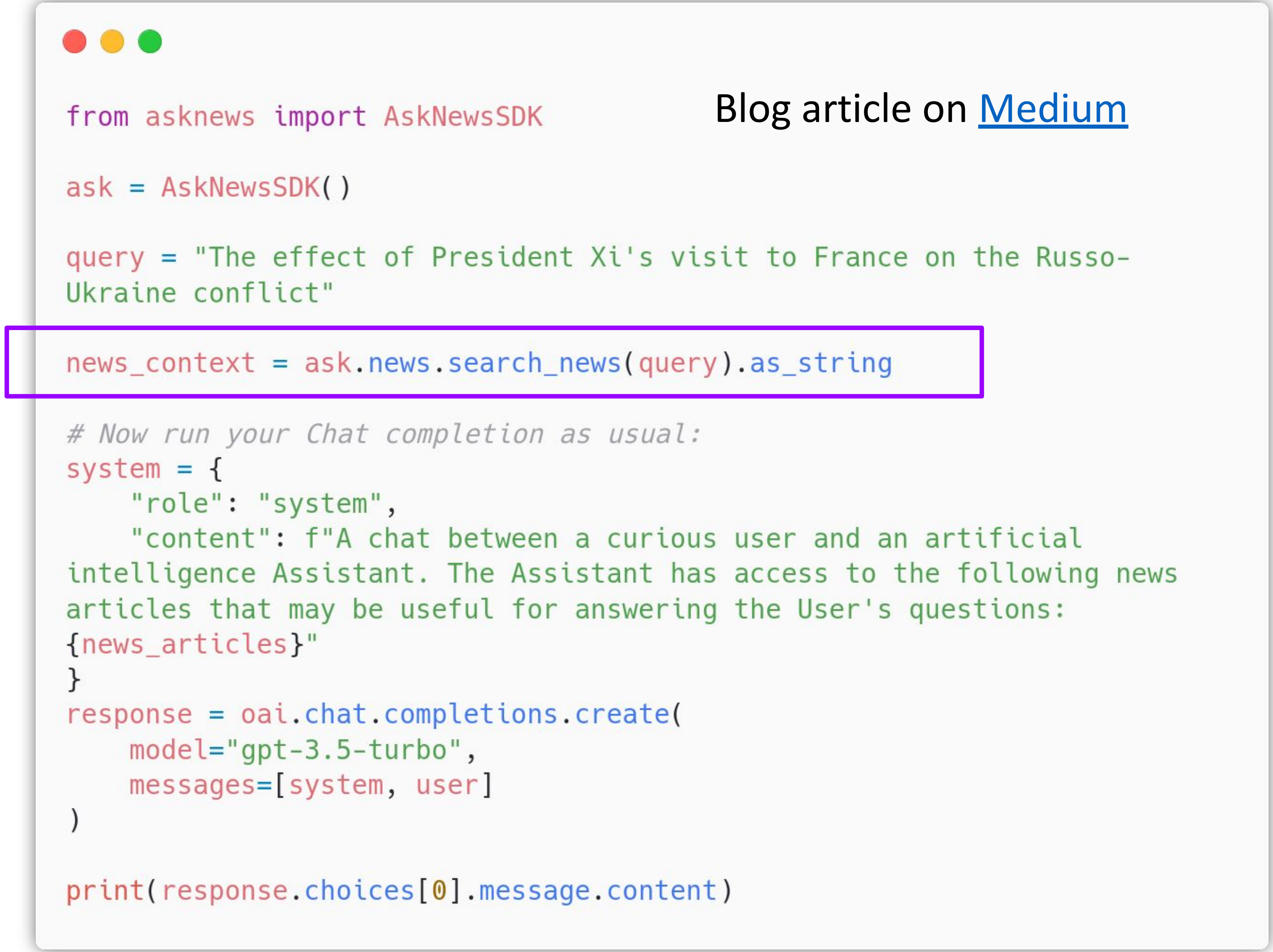
Emergent Methods · Following
4 min read · Feb 23, 2024

Blog article on Medium

Leveraging AskNews for context

Emergent Methods

Outsource your news context to AskNews



Blog article on Medium

```
from asknews import AskNewsSDK

ask = AskNewsSDK()

query = "The effect of President Xi's visit to France on the Russo-
Ukraine conflict"

news_context = ask.news.search_news(query).as_string

# Now run your Chat completion as usual:
system = {
    "role": "system",
    "content": f"A chat between a curious user and an artificial
intelligence Assistant. The Assistant has access to the following news
articles that may be useful for answering the User's questions:
{news_articles}"
}
response = oai.chat.completions.create(
    model="gpt-3.5-turbo",
    messages=[system, user]
)

print(response.choices[0].message.content)
```

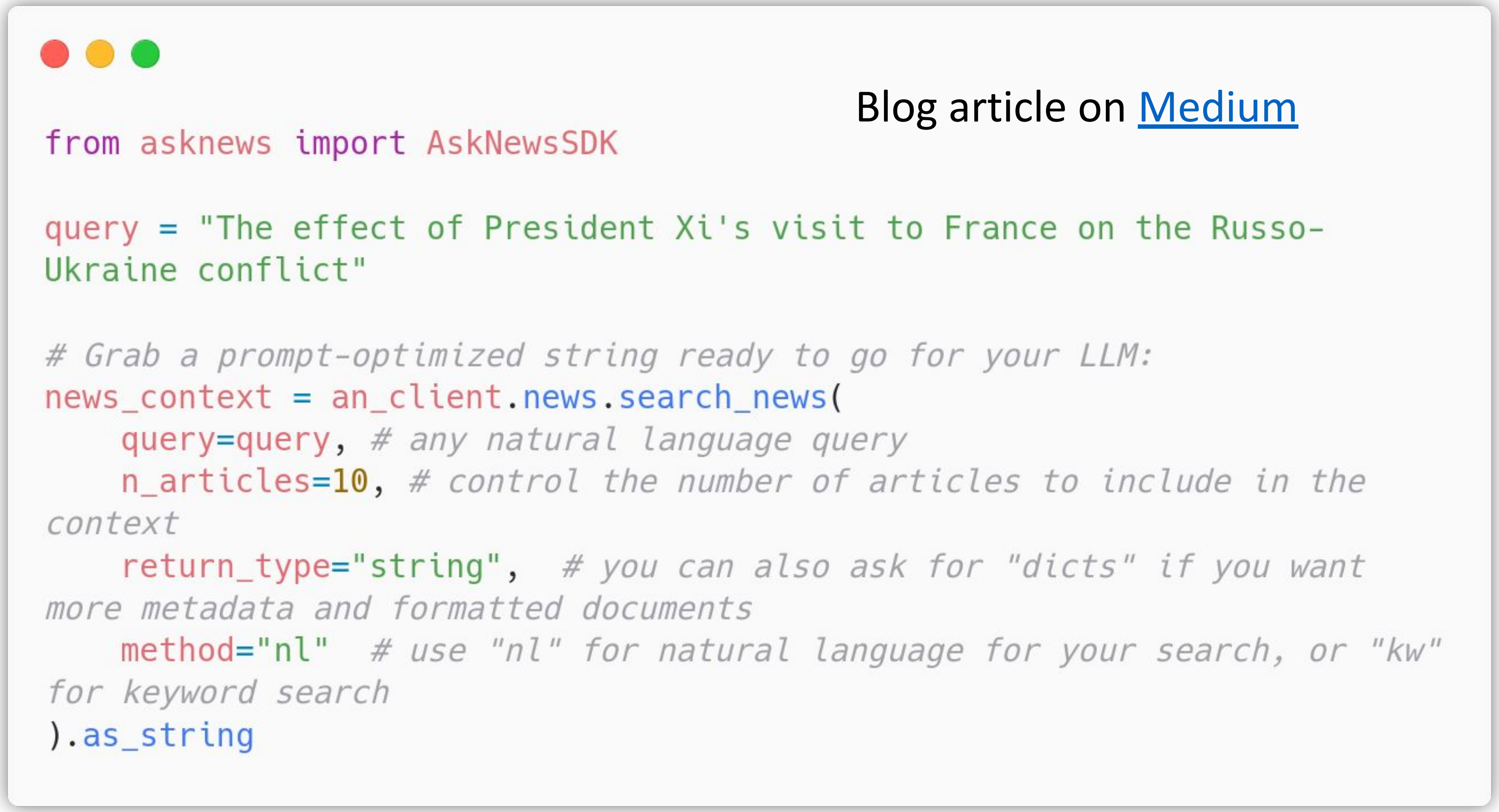
2024-05-07 Unstructured Data Meetup: Multilingual narrative tracking in the news - real-time experiments

2024-05-07 Unstructured Data Meetup: Multilingual narrative tracking in the news - real-time experiments

Emergent Methods
www.emergentmethods.ai

Outsource your news context to AskNews

Blog article on Medium



```python
from asknews import AskNewsSDK

query = "The effect of President Xi's visit to France on the Russo-Ukraine conflict"

# Grab a prompt-optimized string ready to go for your LLM:
news_context = an_client.news.search_news(
    query=query, # any natural language query
    n_articles=10, # control the number of articles to include in the context
    return_type="string",  # you can also ask for "dicts" if you want more metadata and formatted documents
    method="nl"  # use "nl" for natural language for your search, or "kw" for keyword search
).as_string
```

Emergent Methods
www.emergentmethods.ai

Outsource your news context to AskNews

- 🌍 Global coverage - 300k articles per day
- 🏎️ Low-latency (100 ms) - aimed at tight spots in your LLM stack
- 🧺 Stories + clustering, Chat, Finance, Citation control
- 📈 Hot topic following/tracking/filtering
- 📊 Usage tracked - pay for what you use
- 👥 Reddit perspective - include social context
- 🆓 Free news api https://my.asknews.app/plans

Blog article on Medium

AskNews

https://docs.asknews.app

Thanks for your attention!

AskNews

Engineering news context

https://asknews.app

Robert Caulk, PhD & Elin Törnquist, PhD

CEO *Director of Transparency*

Emergent Methods

https://emergentmethods.ai