

Studying Online Communities and Digital Discourse with Communalytic.org



Anatoliy Gruzd

Toronto Metropolitan University, Canada

info@SocialMediaLab.ca

Agenda

1. About at the **Social Media Lab & Social Media Research Landscape**
2. Overview of **Communalytic**
3. **Topic Analysis** with Embeddings

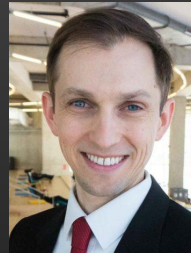
About the Social Media Lab

 [@SocialMediaLab.ca](https://twitter.com/SocialMediaLab.ca)



A Multidisciplinary Research Laboratory at TRSM

The lab's research seeks to advance the public's understanding of the benefits and pitfalls of social media adoption.



Dr. Anatoliy Gruz
Canada Research Chair,
Professor and Co-Director



Philip Mai M.A., J.D.
Co-Director and Senior
Researcher

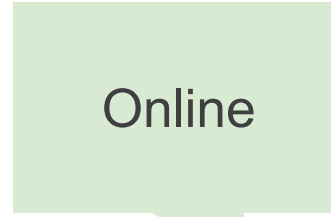


Research Foci

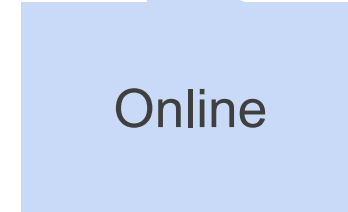


International Conference on Social Media & Society

July 2026 | London, UK



The Social Media Lab hosts a variety of research events throughout the year. These events allow participants to exchange ideas, learn new tools and methods, discover new research and network with peers.



14 Years Running
400+ Authors,
200+ Attendees from
25+ Countries every year

socialmedialab.ca/events/

Recent publication: Exposure to and Belief in Pro-Kremlin Disinformation among Canadians

Soares, F. B., Gruzd, A., & Mai, P. (2023). **Falling for Russian propaganda: Understanding the factors that contribute to belief in pro-Kremlin disinformation on social media.** *Social Media+ Society*, 9(4), 20563051231220330.



<https://doi.org/10.1177/20563051231220330>

Key Findings

1. Exposure to disinformation on social media is positively correlated with belief in it.
2. Individuals who hold right-leaning ideology are more likely to believe in pro-Kremlin claims overall as compared to those who hold mixed or left-leaning views.
3. Those who believe in pro-Kremlin claims are more likely to rely on social media for news about the war than those who do not believe.



Recent publication: Effectiveness of Warning Labels on Posts Containing False Claims about the Russia-Ukraine War

Gruzd, A., Mai, P., & Soares, F. B. (2024).

To Share or Not to Share: Randomized Controlled Study of Misinformation Warning Labels on Social Media. In *Multidisciplinary International Symposium on Disinformation in Open Online Media* (pp. 46-69). Cham: Springer Nature.



https://doi.org/10.1007/978-3-031-71210-4_4

RQ1: RESULTS

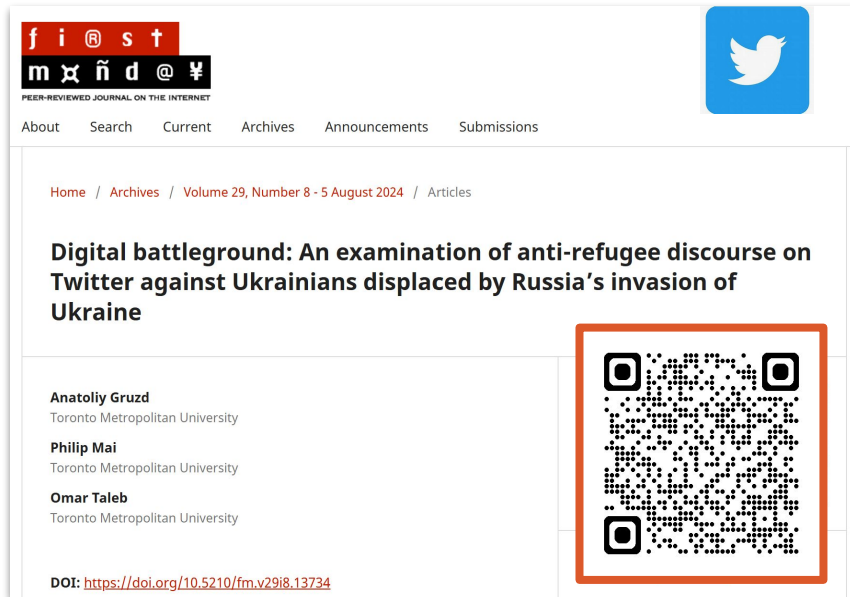
One-way ANOVA/Post hoc Games-Howell test

Do soft moderation interventions commonly used by social media platforms reduce engagement with misinformation?



Both interventions have decreased the mean number of interactions with false claims.

Recent publication: Pro-Kremlin discourse on Twitter



Key Finding

- **53%** of the sample posts aligned with pro-Kremlin narratives against Ukraine.
- By exploiting anti-refugee sentiments and political/cultural fault lines in the West, pro-Kremlin messages contribute to
 - diminishing support for Ukrainian refugees,
 - minimizing the severity of the war, and
 - undermining international support for Ukraine.

PUBLIC REPORTS

Data About Canadians For Canadians

As part of our work, the Lab produces public reports that provide a snapshot of social media usage trends in Canada and a window into Canadians' attitudes and perspectives toward issues such as misinformation, politics, and privacy in the context of social media.

Our public reports are widely used by different stakeholders including parliamentary committees, government departments, journalists, researchers, educators, the business community and the general public.

socialmedialab.ca/public-reports/

Reports

State of Social Media in Canada



Misinformation Research



Social Media and User Privacy



The Influence of Influencers



Social Media and Political Engagement in Canada



Social Media and Anti-Social Behaviour



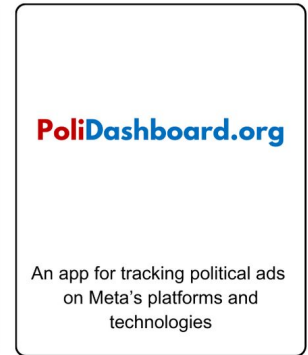
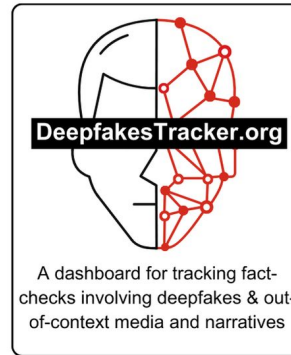
Recent Reports:

- [The State of Anti-Social Behaviour on Social Media 2023](#)
- [The Influence of Influencers in Canada 2023](#)
- [The Reach of Russian Propaganda & Disinformation in Canada](#)
- [State of Social Media in Canada 2022](#)

Research Tools & Apps

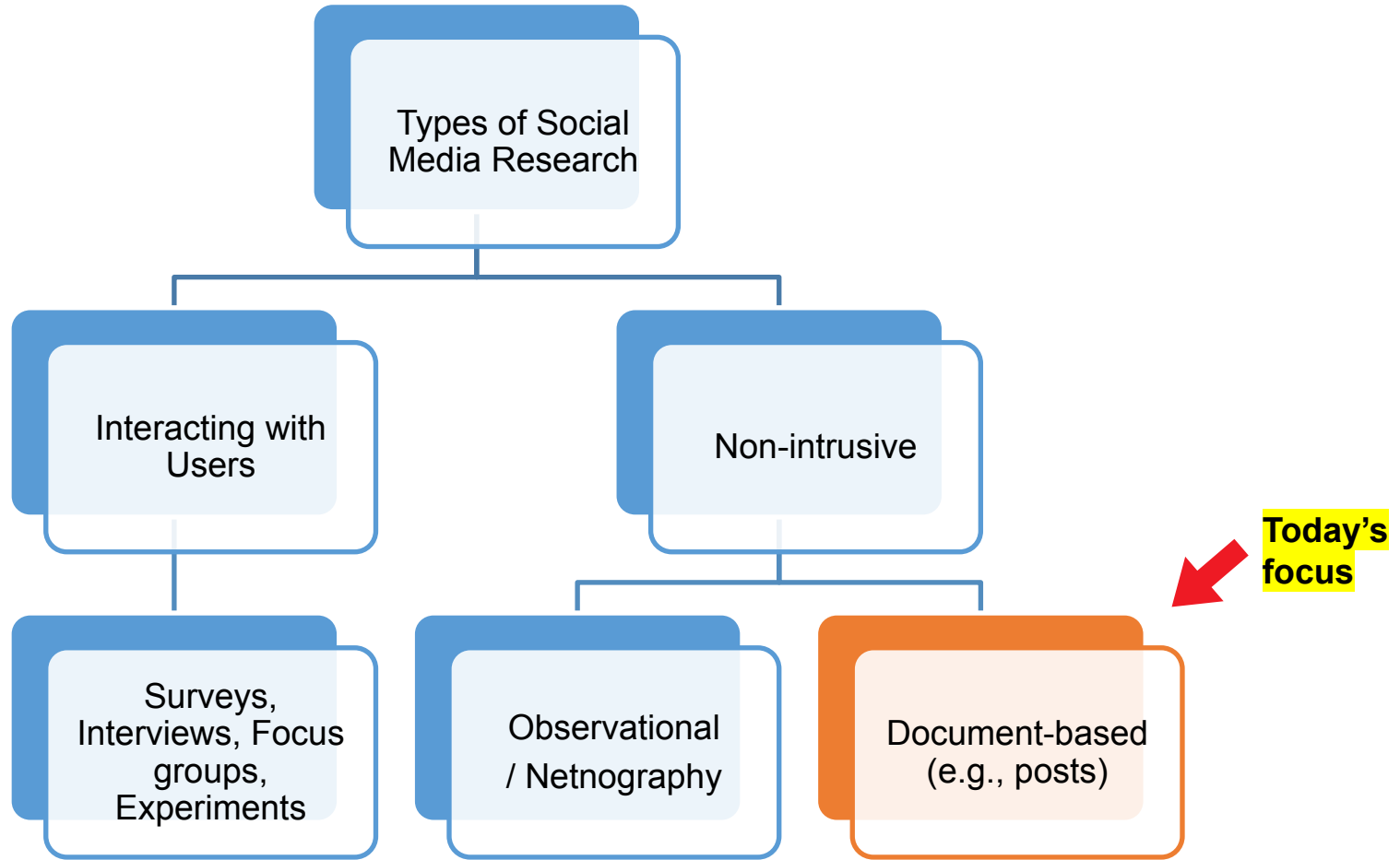
In addition to conducting original research, we also develop tools and dashboards that support social science research and to track misinformation.

Our apps and tools are used by thousands of students, educators, researchers & analysts worldwide each year.

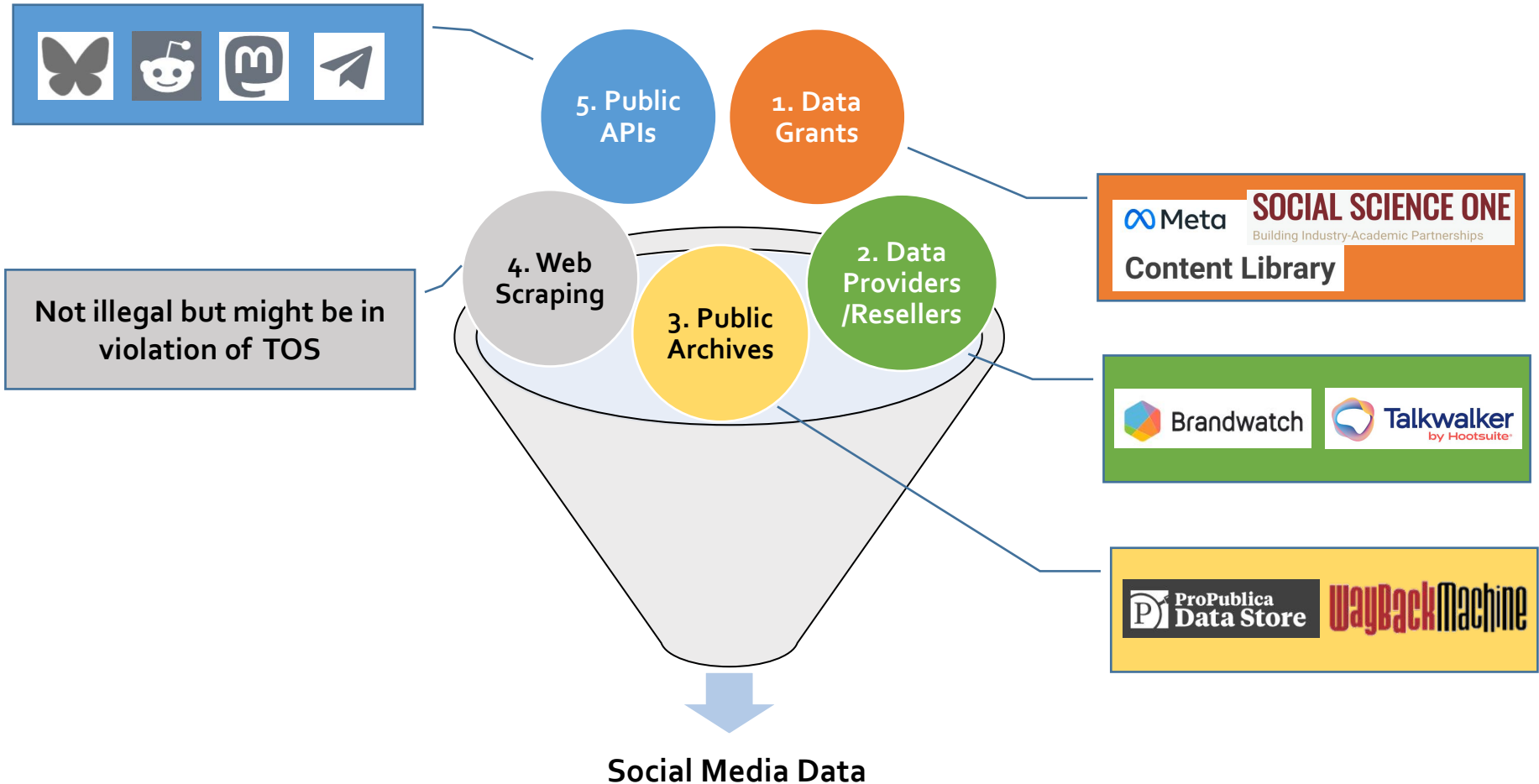


Research Apps, Tools, and Dashboards
socialmedialab.ca/apps

Today's focus



Different Options of Collecting Social Media Data



Social Media Analytics Stages



Ethical Considerations

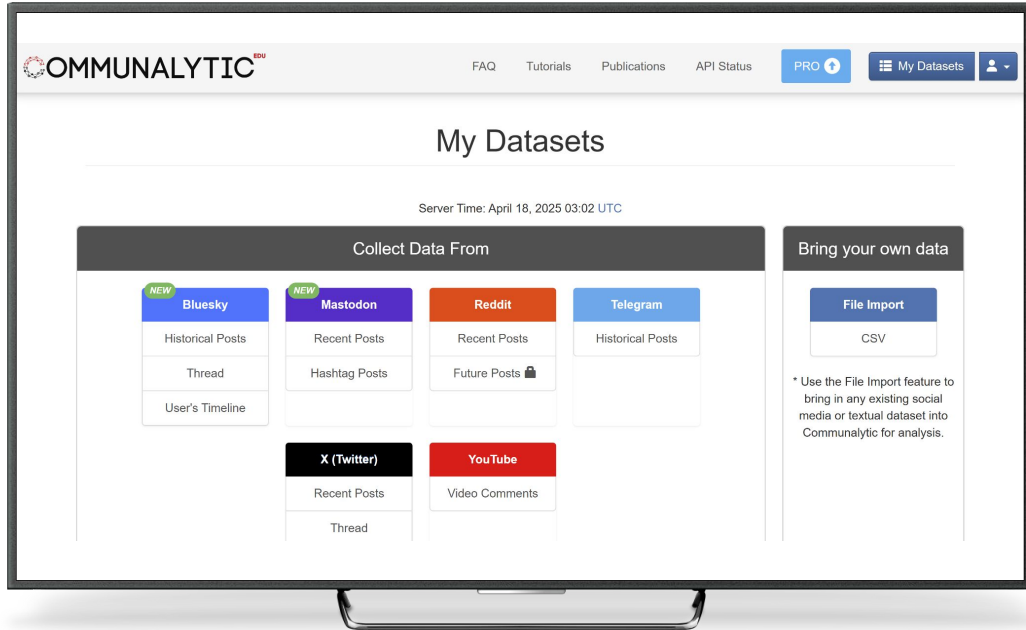
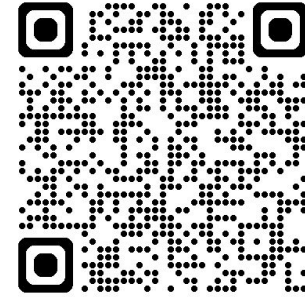
- Social Media Research Ethics → not a “one size fits all” framework
 - Privacy varies across uses, users, and data types
 - Academic researchers > Employer > Financial Companies, Government bodies, Marketers
 - Aggregated > Raw > Social Network Data
 - Research more ‘acceptable’ if:
 - “it’s going to a good cause”
 - “general public good”
 - Social media users care about data quality, accuracy, representation.
 - Special consideration when studying sensitive topics and vulnerable groups (e.g., minors, the deceased, mental health).
- (Social Media Lab) (Golder et al. 2017)

Agenda

1. About at the **Social Media Lab & Social Media Research Landscape**
2. Overview of **Communalytic** 
3. **Topic Analysis** with Embeddings

Research Tool: Communalystic

A research tool developed by TMU's Social Media Lab for studying online communities and discourse



Communalystic can collect, analyze, and visualize public data from social media (e.g. Reddit, Bluesky, Mastodon, Telegram, or CSV files).

<https://communalystic.org>



A no-code computational social science research tool for studying online communities and public discourse. It is designed to provide researchers, journalists and students with essential resources and infrastructure for conducting independent, public-interest research. **Users can bring their own data or use one of Communalytic's various social media data collectors.**

2 Versions of Communalytic

For Teaching & Learning | 30K Records across 3 Datasets

Designed to help students learn about social media data analytics.

EDU Sign Up - It's Free!

➡ EDU Login

For Research | 3M Records across 30 Datasets

Designed for researchers and journalists and is ideal for large-scale research projects.

PRO Sign Up - Free 7-Day Trial

➡ PRO Login

<https://communalytic.org/>



Account Types and Features



Account Types

[EDU] – Free

[PRO] – **New for 2025:**

We now offer more subscription options* and lower prices, starting at just **\$19.99 USD per month.**



Data Storage Cap

[EDU] – Collect $\leq 30K$ records (shared across ≤ 3 datasets)

[PRO] – Collect $\leq 3M$ records (shared across ≤ 30 datasets)



Sign-in Options

[EDU] – via Gmail

[PRO] – via Gmail OR Regular email



Concurrent Data Collection

Collect data from multiple social media platforms at the same time



Collaboration Friendly

Access to various team collaboration features

<https://communalytic.org/>



Data Analyzers

CommunalYTIC also comes with a comprehensive **suite of built-in data analyzers** capable of processing both social and non-social media data. Many of the analyzers are equipped with advanced AI features that assist researchers in conducting complex analyses, offering valuable insights and recommendations as needed. Researchers maintain complete control throughout the process, allowing them to add essential context and expertise and make all final decisions.



Civility Analyzer

Identifies toxic and prosocial interactions in a dataset using the latest machine-learning models



Sentiment Analyzer

Calculates sentiment polarity scores to determine whether text in a dataset expresses a positive, negative or neutral sentiment



Topic Analyzer

Discovers latent topics based on the semantic similarity between posts (aka records) in a dataset using embeddings and multilingual large language models



Network Analyzer

Generates and visualizes various types of networks in a dataset, including communication and link-sharing networks

Today's focus

<https://communalytic.org/>

Use a Gmail email account to sign up or log into Communalytic EDU

For Teaching & Learning | 30K Records across 3 Datasets

Designed to help students learn about social media data analytics.

EDU Sign Up - It's Free!

➡ EDU Login



For Research | 3M Records across 30 Datasets

Designed for researchers and journalists and is ideal for large-scale research projects.

PRO Sign Up - Free 7-Day Trial

➡ PRO Login



COMMUNALYTIC

Social Media Data Collectors

Communalytic is designed to provide researchers and journalists with the resources and infrastructure necessary for conducting independent research in the public interest. Users can **bring their own data** or use one of Communalytic's various **social media data collectors**. *



<https://communalytic.org/>

Collect Data From

NEW

Bluesky

Historical Posts

Thread

User's Timeline

NEW

Mastodon

Recent Posts

Hashtag Posts

Reddit

Recent Posts

Future Posts 

Telegram

Historical Posts

X (Twitter)

Recent Posts

Thread

YouTube

Video Comments

Bring your own data

File Import

CSV

* Use the File Import feature to bring in any existing social media or textual dataset into Communalytic for analysis.

Agenda

1. About at the **Social Media Lab & Social Media Research Landscape**
2. Overview of **Communalytic**
3. **Topic Analysis** with Embeddings 

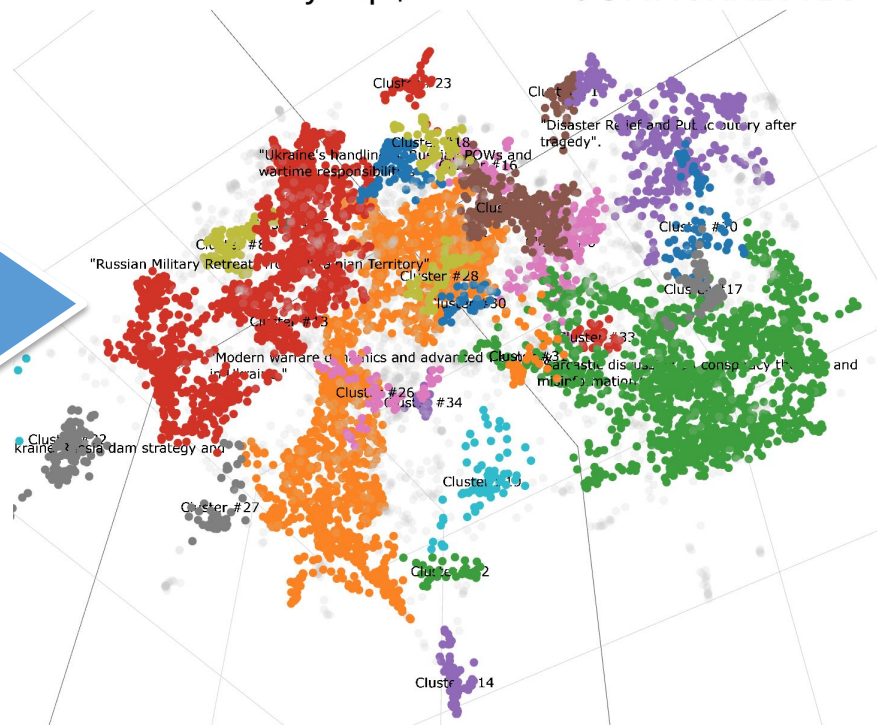
Topic Analysis with Text Embeddings

1. Represent & Project Posts into a Semantic Space

B	C	D	E
date	text	name	username
2022-11-27	And yes, there are just two sexes/genders - a man and a woman, male and female Russia in C RussianEmk		
2022-11-25	It is all about family. Family is a man and a woman and children. https://t.co/Russia in C RussianEmk		
2022-11-27	In Russia you are free to be you & and to protect your children while the Russia in C RussianEmk		
2022-11-27	Twitter said this is hateful content!! Really??? WTF what is the world we are! BcrewA BcrewA		
2022-11-27	#InMemoriam Vladimir Makei https://t.co/1UUbCux2fP https://t.co/H87d4MFA Russia mfa_russia		
2022-11-27	@Amb_Yakovenko from @RussianEmbassy UK. The West must think even Orc Embassy: OrcEmbassy		
2022-11-27	An interview of @BBCNewsnight with Russian ambassador in the UK. "Is your Anton Ger: Gerashchen		
2022-11-25	Hey @russianembassy, does Putin's family (wife, then ex-wife; mistress, no Dr. Ian Gari: Gargner		
2022-11-26	Russian Foreign Minister Sergey Lavrov has said in a documentary aired that TASS tassagency,		
2022-11-25	Заявление Посольства https://t.co/qmKKe2rERT https://t.co/Dzon5DC Russia in C RussianEmk		
2022-11-25	Embassy statement https://t.co/Las3C3ovol https://t.co/XGwydPCg25 Russia in C RussianEmk		
2022-11-27	The @mfa_russia leaders and staff feel deep sorrow over the sudden death MFA Russia: mfa_russia		
2022-11-27	@RussianEmbassyC you guys are fucking bigoted clowns.	Prcurry	Prcurry_ME
2022-11-27	@RussianEmbassyC https://t.co/UkH0Ubo3Op	Professor 7	HiddenOne
2022-11-27	Why does @twitter not allow me to follow @RussianEmbassyC? I click follow no		
2022-11-24	#Zakharova: Russia is making intensive efforts to stabilize the situation in M		
2022-11-27	@RussianEmbassyC https://t.co/NYc33rOrMb		
2022-11-27	imagine the staff of @RussianEmbassyC enjoying the freedom living in Cana: Tibz ACAB: literalD0g		
2022-11-27	@RussianEmbassyC https://t.co/dKZ122sjN	Симми	simmyde12
2022-11-24	Opinion by Maria #Zakharova: The abhorrent Nazi interpretation of eug: MFA Russia: mfa_russia		
2022-11-26	The Embassy was surprised to note the @CanadaFP's publication support in Russia in C RussianEmk		
2019-09-17	The USSR is often accused of invading Poland. Wrong! The Nazis attacked Poland Russia in R Embassyoff		
2022-11-18	For Education His Punishment a sledgehammer to the dome. Circumst Wow that's! BloodGoref		
2022-11-26	RU embassy in Canada @RussianEmbassyC are assholes. #CloseRussianEm WeLoveNA WeLoveNA		
2022-11-26	#OTD in 1812 the Battle of Berezina between retreating Army of #Napole MFA Russia: mfa_russia		
2022-11-25	Today is the Russian Peacekeeper Day – professional day for Russian service! Russia in R Embassyoff		
2022-11-27	Who y'all mad at today? https://t.co/BiancaDel	Bianca Del	TheBiancaD
2022-10-21	"Mark Hamill sent 500 drones to Ukraine in past month", says The Hill. Now, I Russia in R Embassyoff		
2022-11-27	@RussianEmbassyC #Russia #Embassy #CanadaAnd a dog :) https://t.co/ZM2 Freelance 10x09		
2022-11-27	I don't want to change my pin b/c of Twitter f*ckery. But I have now been pr: Evolve Unc: Evolv_unde		
2022-11-26	RUAMBYKZKGTJ The #CSTO Member States urge the global community to abstain MFA Russia: mfa_russia		
2022-10-03	Ukrainian drone caught Russian soldiers during their intimate moment... it ju Russia-Ukr WarFootage		
2022-11-24	#Zakharova: The West is trying its best to turn a blind eye to the terrorist MFA Russia: mfa_russia		
2022-11-18	Ukrainian nationalists have leaked another video of their war crimes and co: Gennady G: Gatilov		

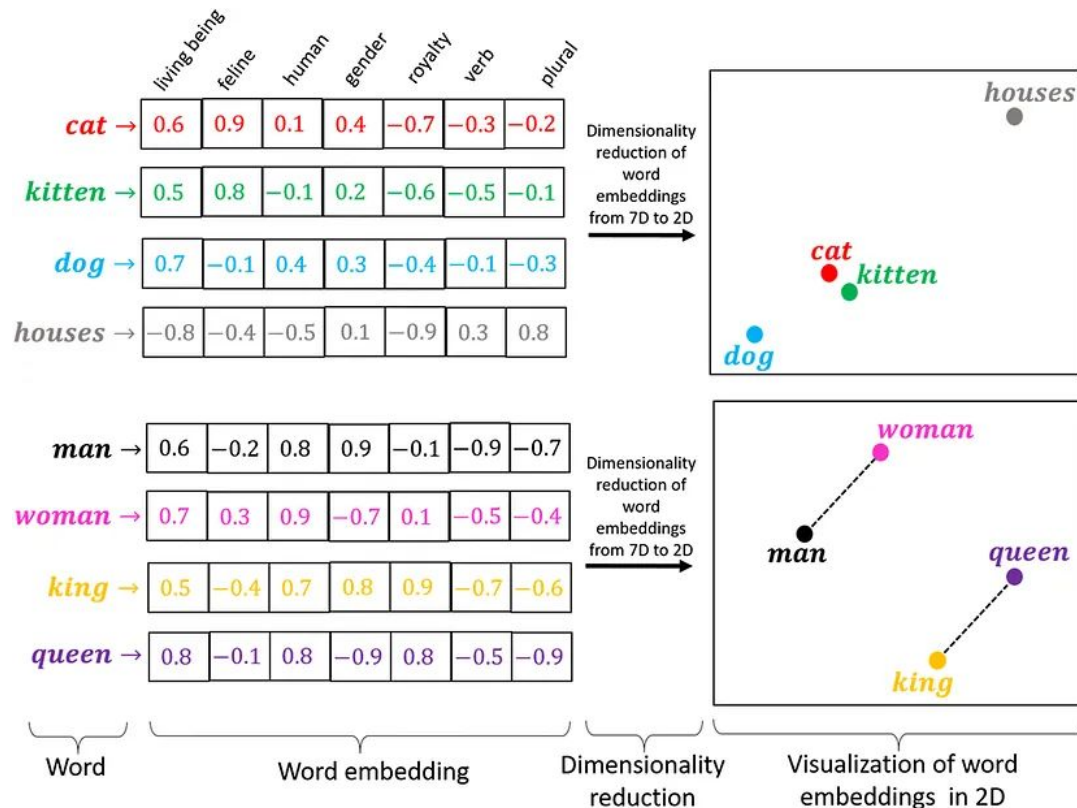
2. Explore Latent Topics

Dataset Name: Ukraine | Platform: Reddit | # Records: 9844
Clustering Algorithm: HDBSCAN (parameters: min cluster size: 30 | min samples: 5 | epsilon: 0.2)
3D Semantic Similarity Map | Made with COMMUNALYTIC



What are text embeddings?

Example of word embeddings:



1. Define “dimensions” and set values per word **or** sentence.
2. Reduce the # of dimensions to visualize them in 2D or 3D.
3. The closer dots (word or sentences) are in the visualization, the closer they are semantically.

What are text embeddings?

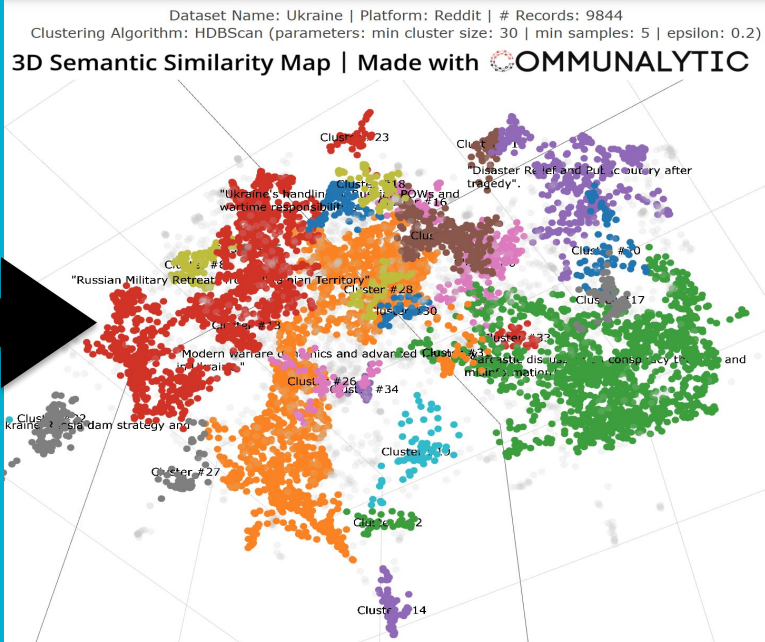
Example of sentence embeddings:

Thank you! →	-0.061	-0.037	-0.002	-0.120	-0.025	-0.007	0.006	-0.004	-0.085	...
Wow! Thanks for your valuable feedback. →	0.011	0.003	0.013	-0.047	-0.032	0.014	-0.082	-0.081	0.000	...
Thank you for the advice. I will try that too :-) →	0.023	-0.092	0.017	-0.038	-0.010	0.026	-0.164	-0.081	-0.028	...

N dimensions (columns)

Main Steps of Communalystic's Topic Analyzer

B	C	D	E
date	text	name	username
2022-11-27	And yes, there are just two sexes/genders - a man and a woman, male and female in C RussianEmb		
2022-11-25	It is all about family. Family is a man and a woman and children. https://t.co/Russia in C RussianEmb		
2022-11-27	In Russia you are free to be you & and to protect your children while the Russia in C RussianEmb		
2022-11-27	Twitter said this is hateful content!! Really??? Wtf what is the world we are l Bwewa ♡ Bwewa A		
2022-11-27	#InMemoriam Vladimir Makai ☹️ https://t.co/jUuBcUx2P https://t.co/H87d4MFA Russia: mfa_		
2022-11-27	👉 @Amb_Yakovenko from @RussianEmbassy UK. The West must think over Embas: OrEmbassy		
2022-11-27	An interview of @BBCNewsnight with Russian ambassador in the UK. "Is your Anton Ger: Gerashchen		
2022-11-25	Hey @russianembassy, does Putin's family (wife, then ex-wife; mistress, n.Cr. Ian Gar: iargner		
2022-11-26	Russian Foreign Minister Sergey Lavrov has said in a documentary aired that I TASS tassageency.		
2022-11-25	👉 Заявление Посольства ☹️ https://t.co/qmXKe2ERT https://t.co/Dzon5Dc Russia in C RussianEmb		
2022-11-25	👉 Embassy statement ☹️ https://t.co/Las3C3ovnl https://t.co/XGwydPCg25 Russia in C RussianEmb		
2022-11-27	■ The @mfa_russia leaders and staff feel deep sorrow over the sudden death MFA Russia: mfa_russia		
2022-11-27	@RussianEmbassyC you guys are fucking bigoted clowns.	Prcury Prcury_ME	
2022-11-27	@RussianEmbassyC https://t.co/UKh0Ubo3Op	Professor z HiddenOne	
2022-11-27	Why does @twitter not allow me to follow @RussianEmbassyC ? I click follow, notABot notabotada		
2022-11-24	👉 #Zakharova: Russia is making intense efforts to stabilize the situation n MFA Russia: mfa_russia		
2022-11-27	@RussianEmbassyC https://t.co/NVc33jOrMb	Ghee gbus WE	
2022-11-27	Imagine the staff of @RussianEmbassyC enjoying the freedom living in Cana: Tibz ACAB: literalD0g		
2022-11-27	@RussianEmbassyC https://t.co/dkZ1225jN	Симми simmyde12	
2022-11-24	#Opinion by Maria #Zakharova: 🤡 The abhorrent Nazi interpretation of eug MFA Russia: mfa_russia		
2022-11-26	The Embassy was surprised to note the @CanadaFP's publication support in Russia in C RussianEmb		
2019-09-17	The USSR is often accused of invading Poland. Wrong! The Nazis attacked Pol Russia in R Embassy		
2022-11-18	👉 For Education 🤡 His Punishment a sledgehammer to the dome. Circumst Wot that: BloodGoref		
2022-11-26	👉 ru embassy in Canada @RussianEmbassyC are assholes. #CloseRussianEm WeLoveNA WeLoveNA		
2022-11-26	👉 OTD in 1812 the Battle of Berezina between retreating Army of #Napole MFA Russia: mfa_russia		
2022-11-25	Today is the Russian Peacekeeper Day – professional day for Russian service: Russia in R Embassy		
2022-11-27	Who 'y'all mad at today? ☹️	Bianca Del TheBiancaD	
2022-10-21	"Mark Hamill sent 500 dollars to Ukraine in past month", says The Hill. Now, I Russia in R Embassy		
2022-11-27	@RussianEmbassyC #Russia #Embassy #CanadaAnd a dog :) https://t.co/ZM2 : Freelance !0x09		
2022-11-27	I don't want to change my pin b/c of Twitter f*ckery. But I have now been pr Evolve UNcEvolve_und		
2022-11-26	RUAMBYKZGTJ The #CSTO Member States urge the global community to absta MFA Russia: mfa_russia		
2022-10-03	Ukrainian drone caught Russian soldiers during their intimate moment... it ju Russia-Ukr WarFootage		
2022-11-24	👉 #Zakharova: The West is trying its best to turn a blind eye to the terrorist MFA Russia: mfa_russia		
2022-11-18	* Ukrainian nationalists have leaked another video of their war crimes and co Gennady GGatitlov		



Visualizing Posts in a 3D Semantic Space

1. Pre-processing posts

- Excluding empty and posts shorter than 3 words;
- Posts over 300 words are truncated.

Thank you!

Wow! Thanks for your valuable feedback.

Thank you for the advice.
I will try that too :-)

2. Representing posts as vector embeddings

- Representing posts as vector embeddings with a multilingual embedding model

-0.061	-0.037	-0.002	...
0.011	0.003	0.013	...
0.023	-0.092	0.017	...

N dimensions

Visualizing Posts in a 3D Semantic Space

1. Pre-processing posts

- Excluding empty and posts shorter than 3 words;
- Posts over 300 words are truncated.

Thank you!

Wow! Thanks for your valuable feedback.

Thank you for the advice.
I will try that too :-)

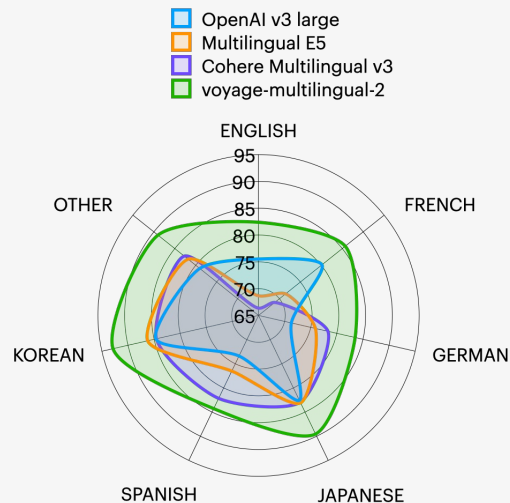
2. Representing posts as vector embeddings

- Representing posts as vector embeddings with a multilingual embedding model

-0.061	-0.037	-0.002	...
0.011	0.003	0.013	...
0.023	-0.092	0.017	...

N dimensions

Communalytic uses VoyageAI API to transform text into embeddings with 1024 dimensions



Supported languages (27+)

- Arabic
- Bengali
- Czech
- Danish
- Dutch
- English
- French
- Georgian
- German
- Greek
- Hungarian
- Italian
- Japanese
- Korean
- Kurdish
- Norwegian
- Persian
- Polish
- Portuguese
- Russian
- Slovak
- Spanish
- Swedish
- Thai
- Turkish
- Urdu
- Vietnamese

Visualizing Posts in a 3D Semantic Space

1. Pre-processing posts

- Excluding empty and posts shorter than 3 words;
- Posts over 300 words are truncated.

Thank you!

Wow! Thanks for your valuable feedback.

Thank you for the advice.
I will try that too :-)

2. Representing posts as vector embeddings

- Representing posts as vector embeddings with a multilingual embedding model

-0.061	-0.037	-0.002	...
0.011	0.003	0.013	...
0.023	-0.092	0.017	...

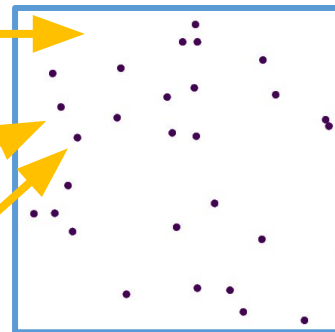
N dimensions

3. Projecting embeddings into a 3D space

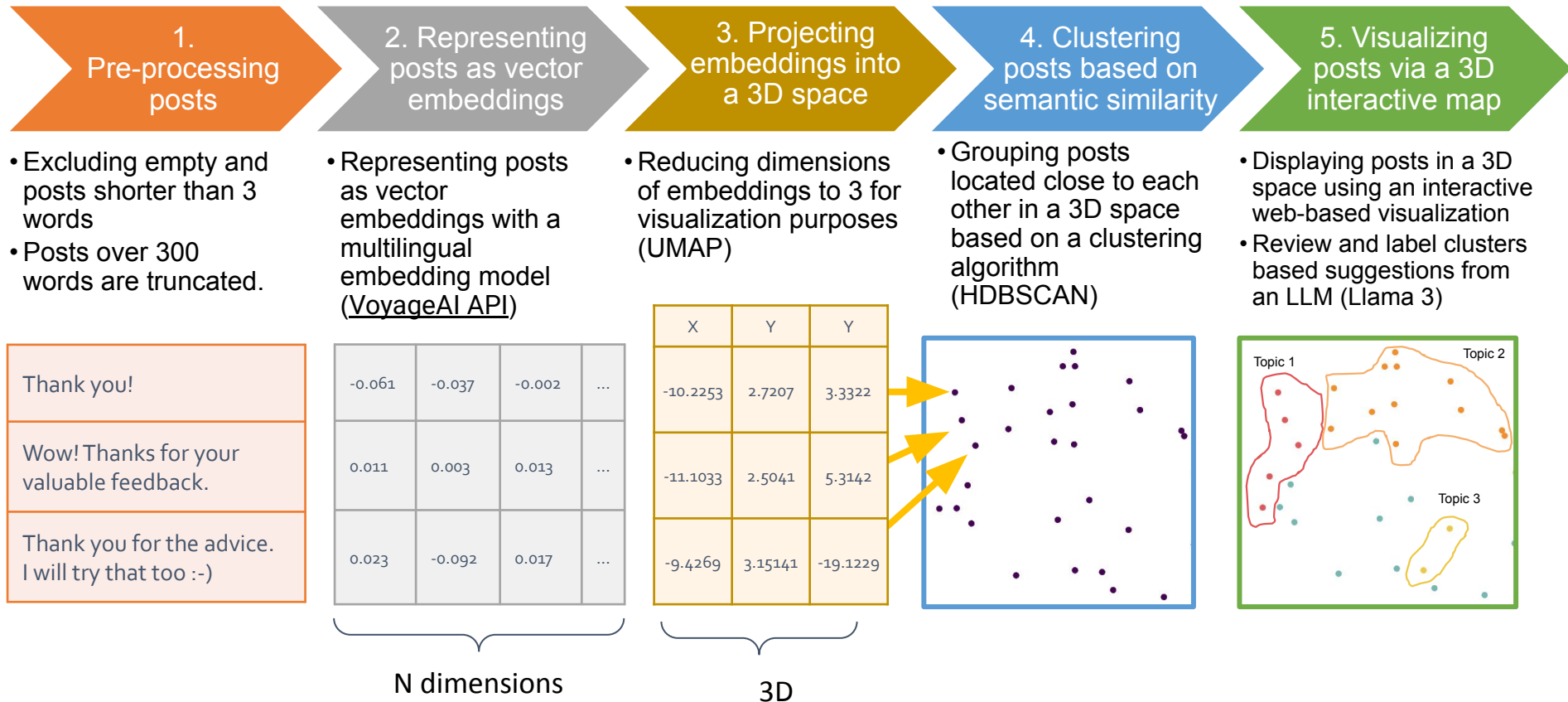
- Reducing dimensions of embeddings to 3 for visualization purposes

X	Y	Y
-10.2253	2.7207	3.3322
-11.1033	2.5041	5.3142
-9.4269	3.15141	-19.1229

3D

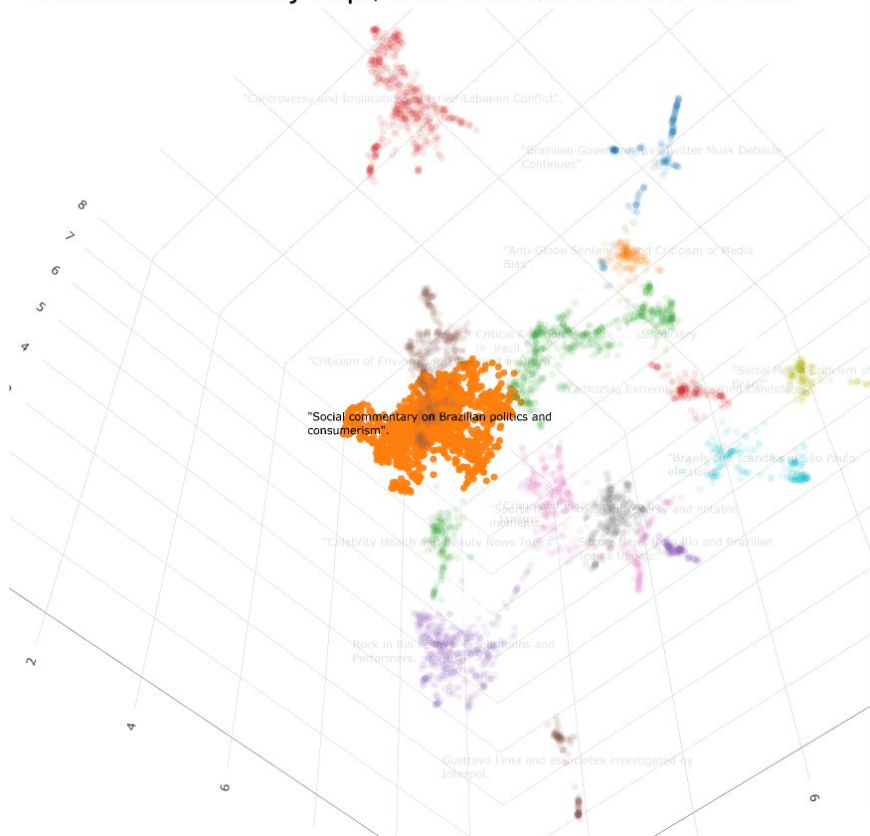


Visualizing Posts in a 3D Semantic Space



Dataset Name: sample | Platform: Bluesky | # Records: 5416
Clustering Algorithm: HDBScan (parameters: min cluster size: 100 | min samples: 1 | epsilon: 0.2)

3D Semantic Similarity Map | Made with COMMUNALYTIC



Navigation Instructions ▾

Examine & Label Clusters ▲

☐ Hide Labels

Select Cluster (Total 16):
1: "Social commentary on Brazilian | ▾

Label Cluster:
"Social commentary on Brazilian politics and consumerism".

Save Cluster Label

llama-3.1-8b-instruct ▾ ?

Suggest a Label (28/30)

Posts in the Selected Cluster

3 / 1000 ⬅ ➡

Date: 2024-09-01T00:00:00
Text: Cães compreendem as palavras emitidas por botões de som de vídeos virais, diz estudo
<https://oglobo.globo...>

Adjust Clustering Parameters ▾

Semi-Automated Labeling of Clusters

Auto-Labeling with a Large Language Model

- Automatically labels each cluster
- Synthesizes short topical descriptions for each topic/cluster

Hands-on Exercise:

Posts mentioning “**Azerbaijan**” on Bluesky, following the crash of Azerbaijan Airlines Flight 8243 on December 25, 2024

Step 1: Download a zipped CSV file: <https://bit.ly/hicss25-dataset> and import it into CommunalYTIC



Collect Data From

NEW

Bluesky

Historical Posts

Thread

User's Timeline

NEW


Mastodon

Recent Posts

Hashtag Posts

Reddit

Recent Posts

Future Posts 

Telegram

Historical Posts

X (Twitter)

Recent Posts

Thread

YouTube

Video Comments

Bring your own data

File Import

CSV

* Use the File Import feature to bring in any existing social media or textual dataset into CommunalYTIC for analysis.

Azerbaijan Airlines Flight 8243



4K-AZ65, the aircraft involved in the accident

Occurrence

Date	25 December 2024
Summary	Crashed on approach during emergency landing after control failure, accidental shutdown; under investigation

Practice Dataset:

<https://bit.ly/hicss25-dataset>














Step 2: Click on the Topic Analysis icon

My Datasets [Shared With Me](#)

Total records in the account: 1077 / 30000 ?

4%

Dataset Name	Records	Civility Analyzer	Sentiment Analyzer	Topic Analyzer	Network Analyzer	Download	Share	Delete
 hicss dataset 	 Records: 1077			 				

Step 3: Click Visualize Embeddings

Clustering Parameters ☒

Select a Clustering Algorithm:

Fast HDBScan

Epsilon: ?

0.1

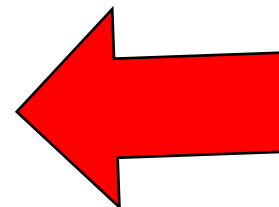
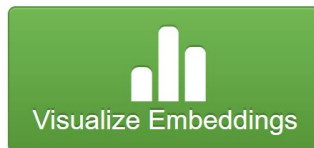
Minimum Cluster Size: ?

10

Minimum Sample Size: ?

10

☐ Check to be notified when visualization is ready.



Step 4: Wait

Embeddings Analysis in Progress...

(You may close this window and visit it later to review/download the results)

45%

Progress: 4699 / 5171 posts analyzed


Check progress in **57** sec

Cancel Analysis

Step 5: Click Open Visualization

Topic Analysis Module (Experimental)

The Topic Analysis Module automatically identifies and groups together social media posts that are semantically similar and can be used to spot latent topics in a dataset (i.e., abstract topics that may not be directly observable from just reading the posts). The module transforms human-readable textual data such as social media posts into computer-readable vectors of numbers known as embeddings. To get start, please review [this tutorial](#).

Learn more... 



Open
Visualization



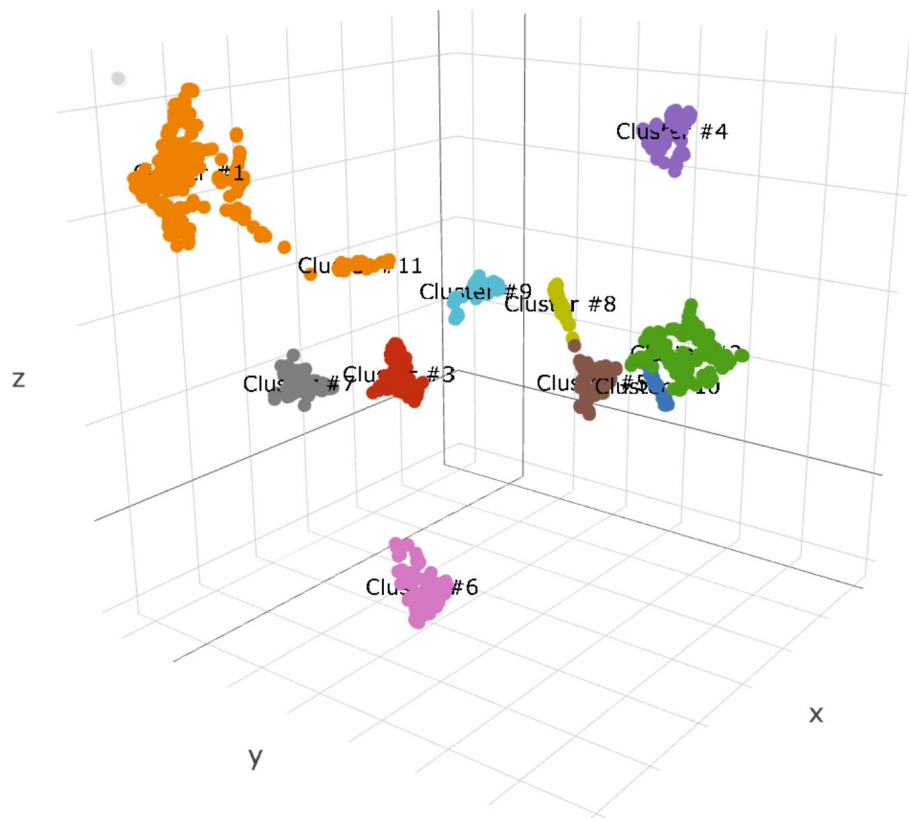
Change
Clustering Parameters




Download
Embeddings & Clusters



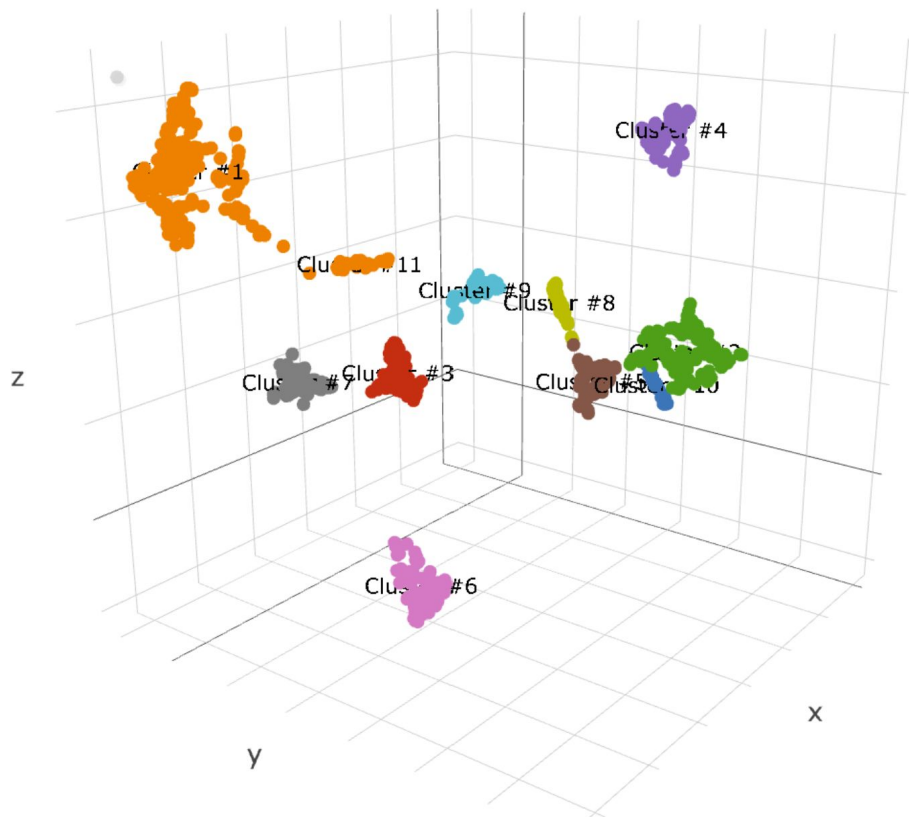
Step 6: Explore the 3D Semantic Similarity Map



Navigation Instructions ▼

- Use the **drop down menu** in the **"Examine & Label Clusters"** panel to **select** and zoom in on a specific cluster.
- **Click the**  **icon** at the top right corner to **see where the selected cluster is located** within the 3D Map.
- **Left click** mouse and move to **rotate the camera**.
- **Right click** mouse and move to **pan the camera**.
- **Scroll** mouse wheel to **zoom in and out**.
- **Hover** on dot to **preview post**.

Step 7: Review & Label Clusters



Examine & Label Clusters ▾

☐ Hide Labels

Select Cluster (Total 14):
2: Cluster #2 ▾

Label Cluster:

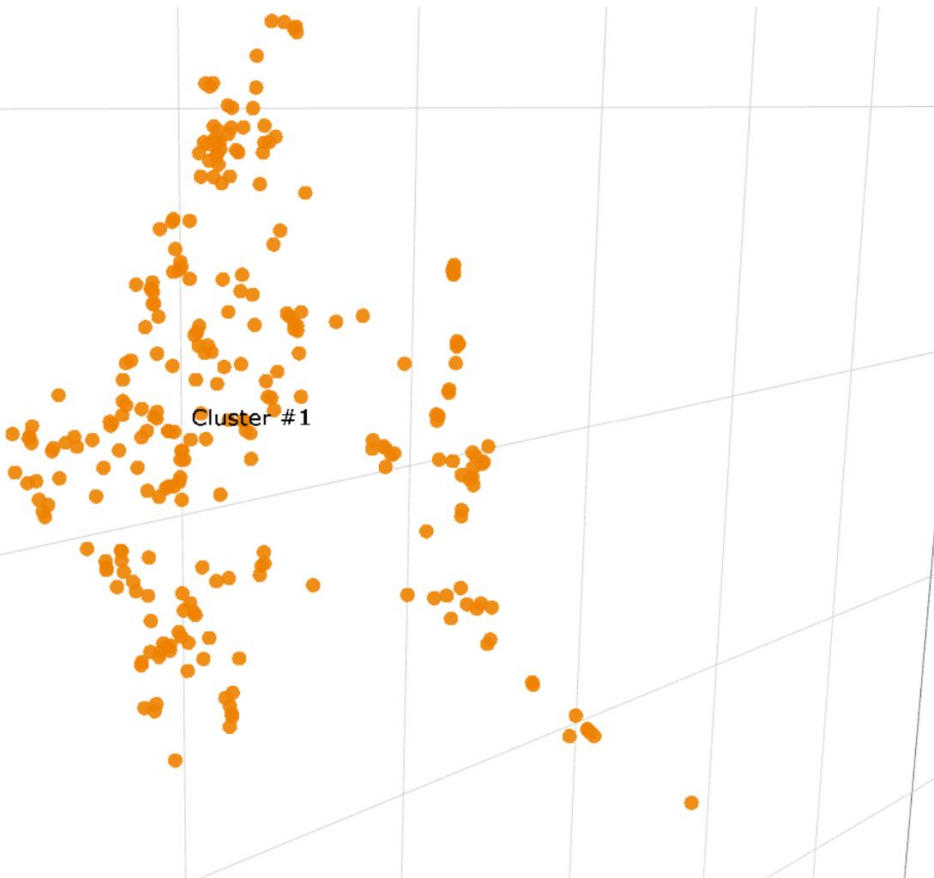
Save Cluster Label

llama-3.1-8b-instruct ▾ ?

Suggest a Label (26/30)

Posts in the Selected Cluster
4 / 533 < >
Date: 2024-01-02T16:10:06.751
Text: Ana Ofelia Murguía, one of Mexico's most acclaimed actresses, was born on January 2, 1941, in Mexico City.

Step 7: Review & Label Clusters



Navigation Instructions ▼

Examine & Label Clusters ▲

☐ Hide Labels

Select Cluster (Total 11):
1: Cluster #1 ▼

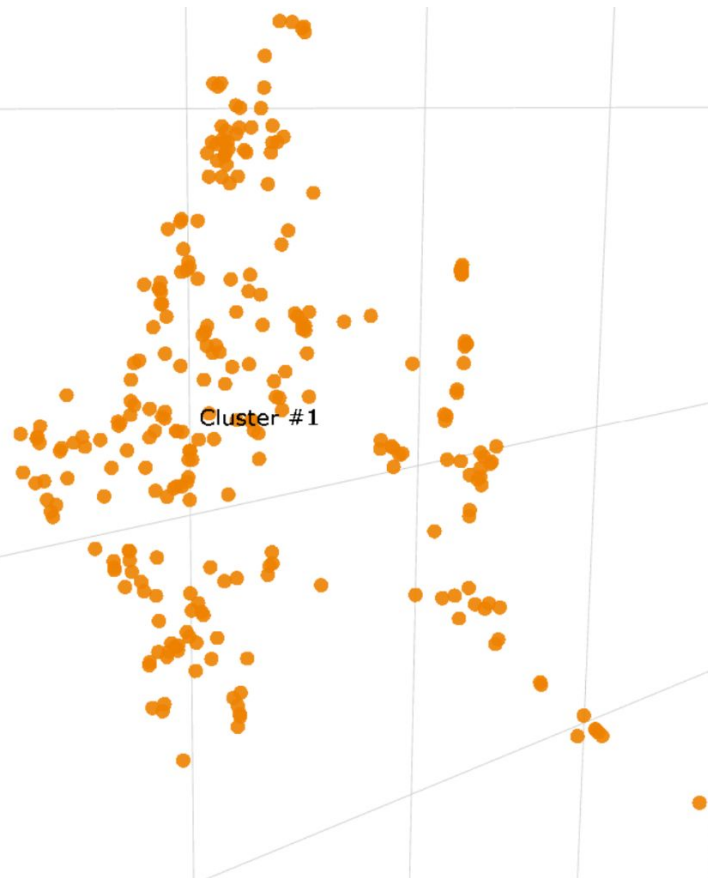
Label Cluster:
"Russian President Putin Apologizes for Plane Crash"

Note: AI-generated label may be inaccurate or unclear; we suggest to review a sample of posts (5-10%) from this cluster and modify the label as needed.

Save Cluster Label

llama-3.2-3b ▼ ?

Step 7: Review & Label Clusters



Navigation Instructions ▼

Examine & Label Clusters ▲

☐ Hide Labels

Select Cluster (Total 11):

1: "Russian President Putin Apologiz" ▼

Label Cluster:

Enter a Cluster Label...

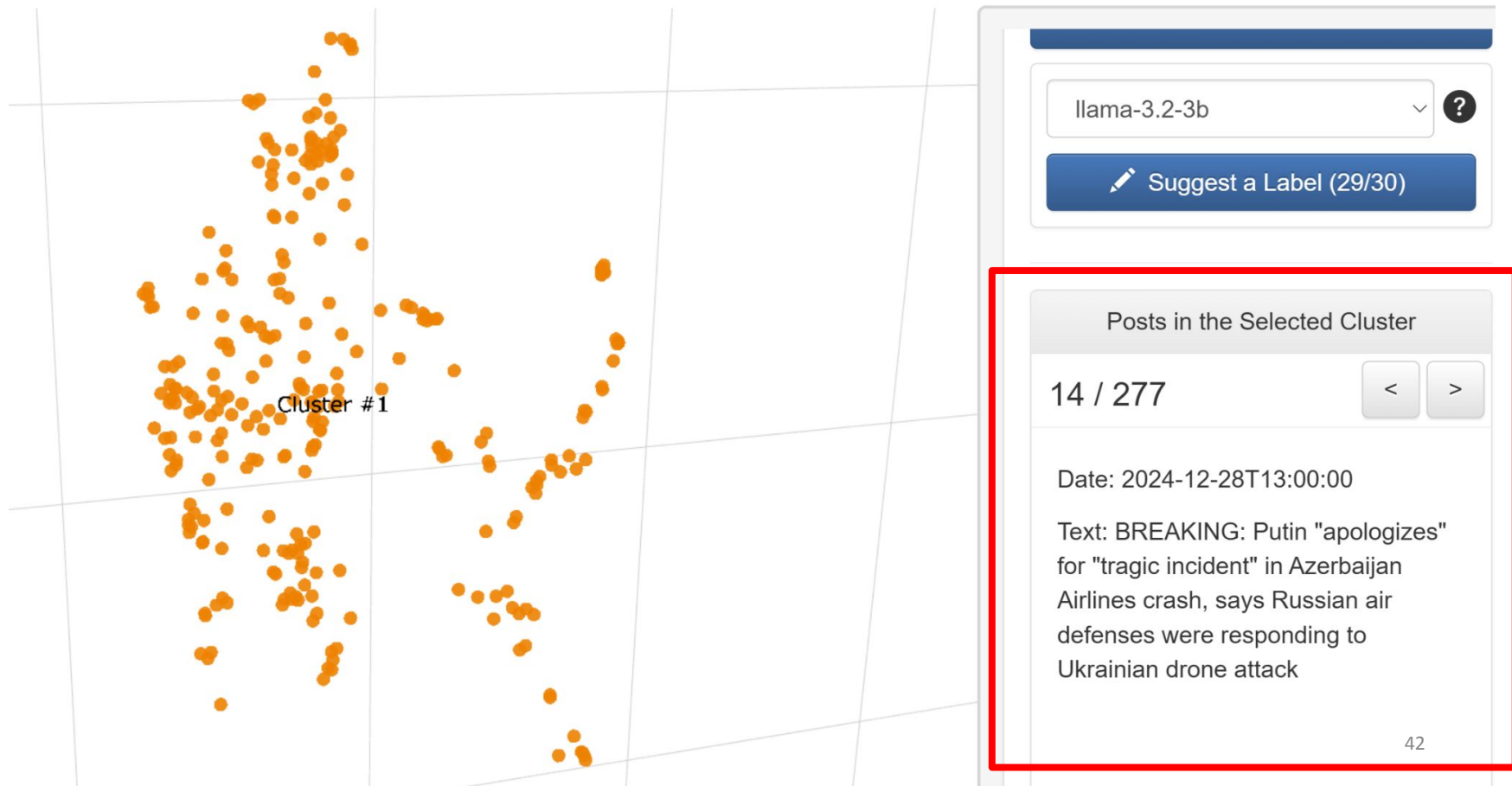
Note: AI-generated label may be inaccurate or unclear; we suggest to review a sample of posts (5-10%) from this cluster and modify the label as needed.

Save Cluster Label

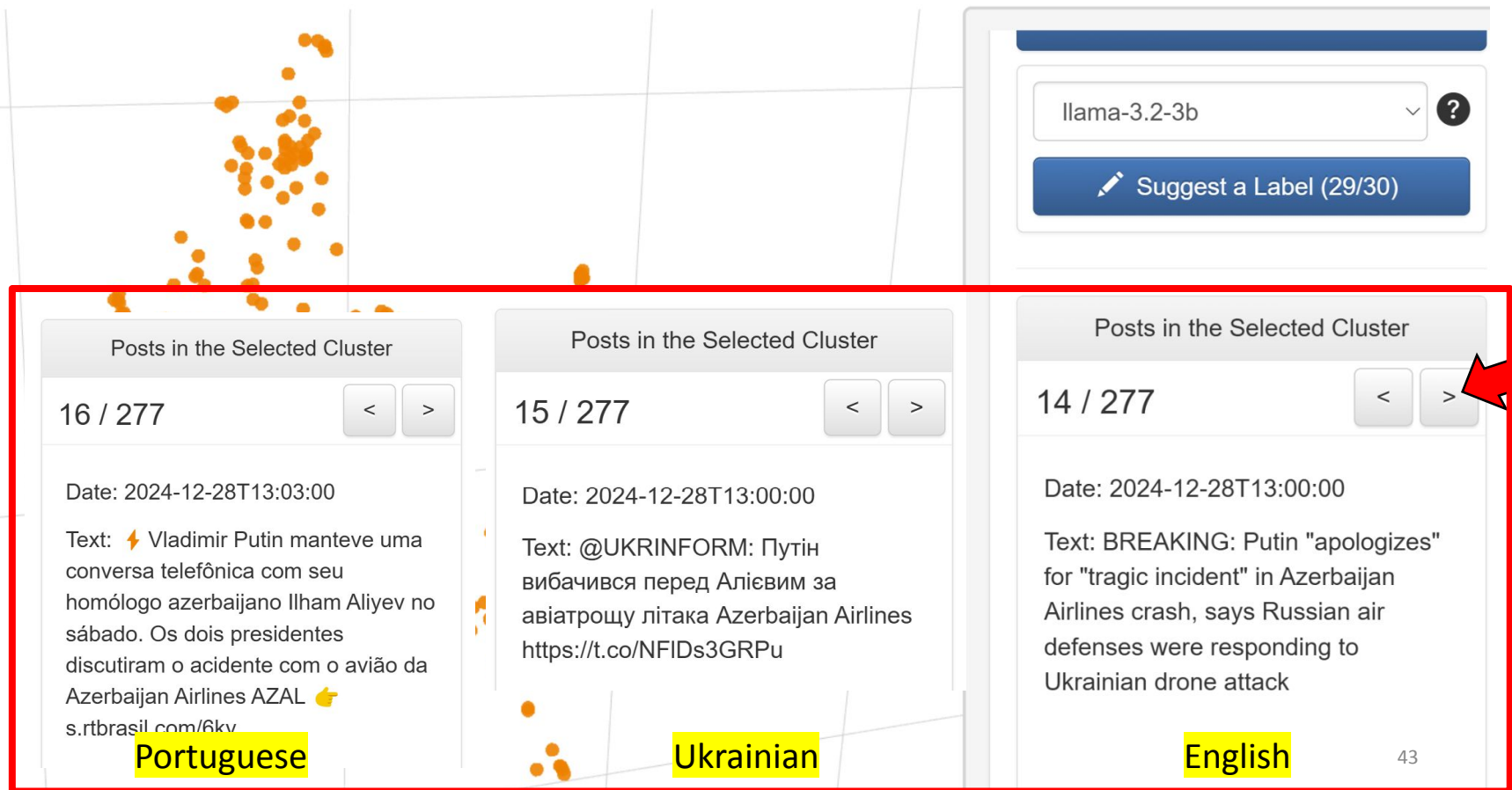
✓ Saved

llama-3.2-3b ▼ ?


Step 8: Cross-validate each AI-suggested label



Step 8: Cross-validate each AI-suggested label



llama-3.2-3b ?

 Suggest a Label (29/30)

Posts in the Selected Cluster

16 / 277 < >

Date: 2024-12-28T13:03:00

Text: ⚡ Vladimir Putin manteve uma conversa telefônica com seu homólogo azerbaijano Ilham Aliyev no sábado. Os dois presidentes discutiram o acidente com o avião da Azerbaijan Airlines AZAL 👉
s.rtbrasil.com/6kv

Portuguese

Posts in the Selected Cluster

15 / 277 < >

Date: 2024-12-28T13:00:00

Text: @UKRINFORM: Путін вибачився перед Алієвим за авіатроцю літака Azerbaijan Airlines
<https://t.co/NFIDs3GRPu>

Ukrainian

Posts in the Selected Cluster

14 / 277 < >

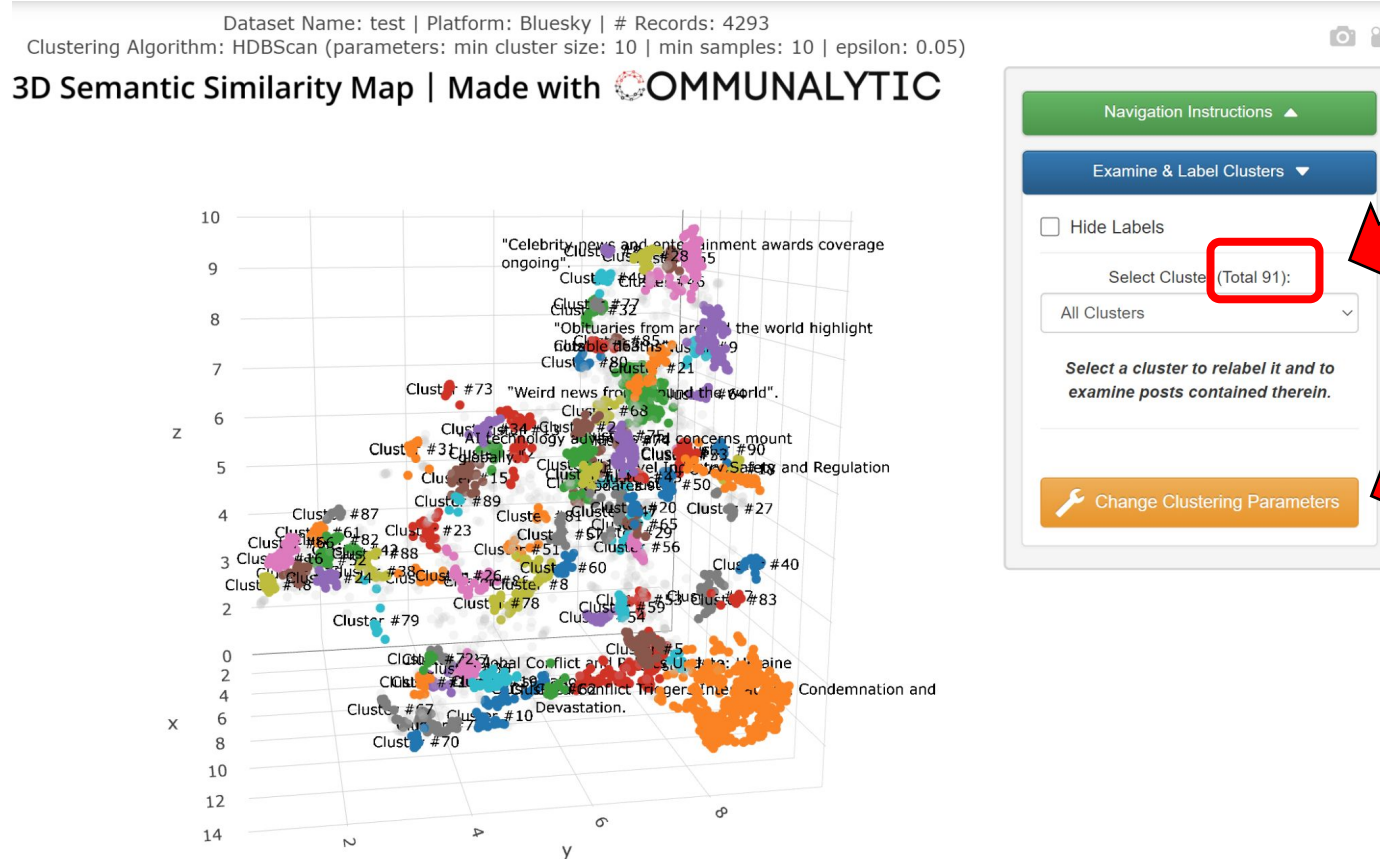
Date: 2024-12-28T13:00:00

Text: BREAKING: Putin "apologizes" for "tragic incident" in Azerbaijan Airlines crash, says Russian air defenses were responding to Ukrainian drone attack

English

43

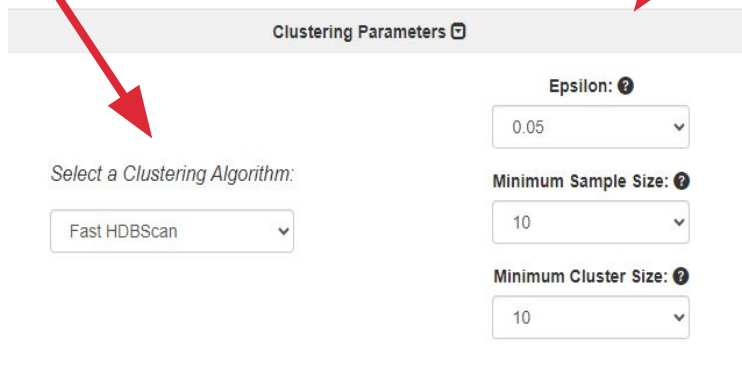
Step 9: Change clustering parameters to reduce/increase the number of clusters (Optional)



#Clusters

Supported Clustering Algorithms

- **Fast HDBScan** (the default) is best for data with varying densities and outliers but requires parameter tuning.
- **KMeans** is efficient and easy to use but assumes spherical clusters and is sensitive to outliers.
- **Gaussian Mixture** offers flexibility in cluster shapes with probabilistic memberships but assumes Gaussian distributions.



The image shows a 'Clustering Parameters' dialog box. It has a title bar 'Clustering Parameters' with a close button. Below the title bar, there is a section 'Select a Clustering Algorithm:' with a dropdown menu showing 'Fast HDBScan'. To the right of this section, there are three parameters, each with a label, a help icon, and a dropdown menu: 'Epsilon: ?' with a value of '0.05', 'Minimum Sample Size: ?' with a value of '10', and 'Minimum Cluster Size: ?' with a value of '10'. Two red arrows point from the text 'Fast HDBScan' in the first bullet point of the 'Supported Clustering Algorithms' section to the 'Fast HDBScan' dropdown and the 'Epsilon' parameter dropdown respectively.

Clustering Parameters

Select a Clustering Algorithm:

Fast HDBScan

Epsilon: ?

0.05

Minimum Sample Size: ?

10

Minimum Cluster Size: ?

10

HDBScan Parameters

- **Epsilon** controls the distance for neighborhood consideration, impacting how close points need to be to form a cluster.
- **Minimum Sample Size** determines the density requirement for a point to be considered a core point and thus part of a cluster.
- **Minimum Cluster Size** sets a threshold for the minimum number of posts required to form a cluster.

Adjusting Clustering Parameters based on the Davies-Bouldin Index (DBI)

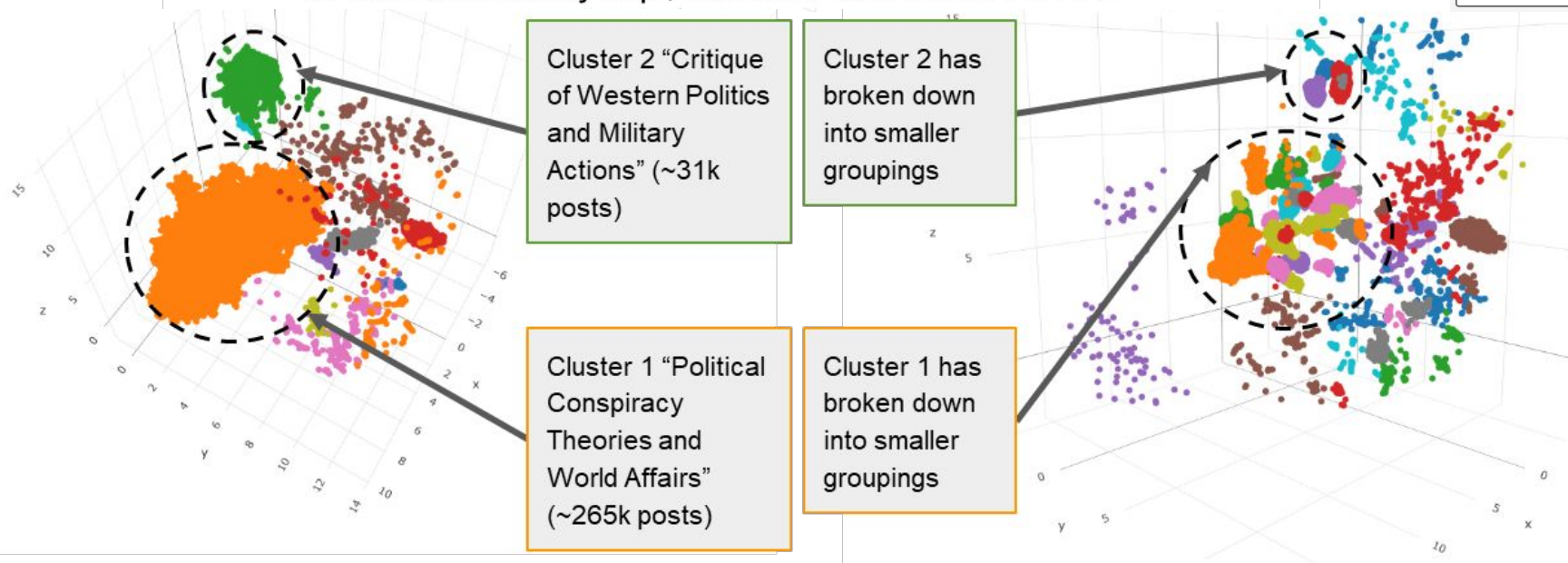
Dataset Info	
Dataset Name:	Ukraine - Feb 21
Platform:	Twitter
Records:	640458
Dataset Parameters	
Search Criteria:	Ukraine
Collection started:	2022-02-21 00:00
Collection ended:	2022-02-22 00:00

Dataset Name: Ukraine - Feb 21 | Platform: Twitter | # Records: 525020 | # Clusters: 16 | # Outliers: 162646

Clustering Algorithm: HDBSCAN (parameters: min cluster size: 1000)

Clustering Quality: Davies-Bouldin Index=0.612 (Lower=Better)

3D Semantic Similarity Map | Made with COMMUNALYTIC

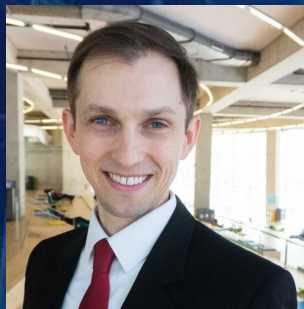


Minimum Cluster Size = 1000

Minimum Cluster Size = 500

Related tutorial: <https://communalytic.org/docs/adjusting-clustering-parameters/>

Studying Online Communities and Digital Discourse with Communalytic.org



Anatoliy Gruzd

Toronto Metropolitan University, Canada

info@SocialMediaLab.ca