Marija Mitrović Dankulov[1], Aleksandar Tomašević[2], Slobodan Maletić[3], Ana Vranić[1], Miroslav Anđelković[3], Darja Cvetković[1], Boris Stupovski[1], Dušan Vudragović[1], Sara Major[2], Aleksandar Bogojević[1]

[1]Institute of Physics Belgrade, Belgrade, Serbia
[2]Faculty of Philosophy, University of Novi Sad, Serbia
[3]"Vinča" Institute of Nuclear Sciences, National Institute of the Republic of Serbia, University of Belgrade, Belgrade, Serbia

MADOC : Multi-Platform Aggregated Dataset of Online Communities

Platforms

Bluesky Koo Reddit Voat

18.9 M posts

236 M comments

23.1 M users

1. Community / Topic alignment using LDA

funny gaming gifs pics
videos technology
fatpeoplehate GreatAwakening
MillionDolarExtreme
CringeAnarchy KokatuInAction
MensRights

2. Sentiment

- VADER
- TextBlob
- Toxigen RoBERTa

3. Applications

- Community dynamics
- Users migrations between platforms
- Toxic behavior

Standardized Data & Access

download dataset

zenodo

import packages

```
pip install pymadoc
devtools::install_github("atomashevic/rMADOC")
```

1. Topic alignment

- LDA topic modeling on 12 Reddit-Voat community pairs (6 general, 6 controversial)
- Balanced sampling strategy: 20,000 documents per topic model
- Top 20 keywords per topic, filtered to retain least frequent third of English words
- Cross-platform filtering: basic (1 keyword) and strict ($\geq$ 2 keywords) criteria
- Results: 910K Bluesky, 3.1M Koo posts (basic); 114K Bluesky, 250K Koo posts (strict)

2. Sentiment

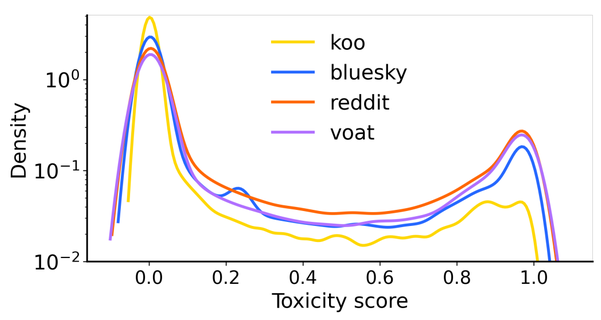
- VADER:
  Social media-aware sentiment (-1 to 1)
- TextBlob:
  Polarity (-1 to 1) and subjectivity (0 to 1) scores
- ToxiGen RoBERTa:
  Implicit toxicity detection (0 to 1)
- Bluesky most positive (0.088), Voat least positive (0.011) mean sentiment

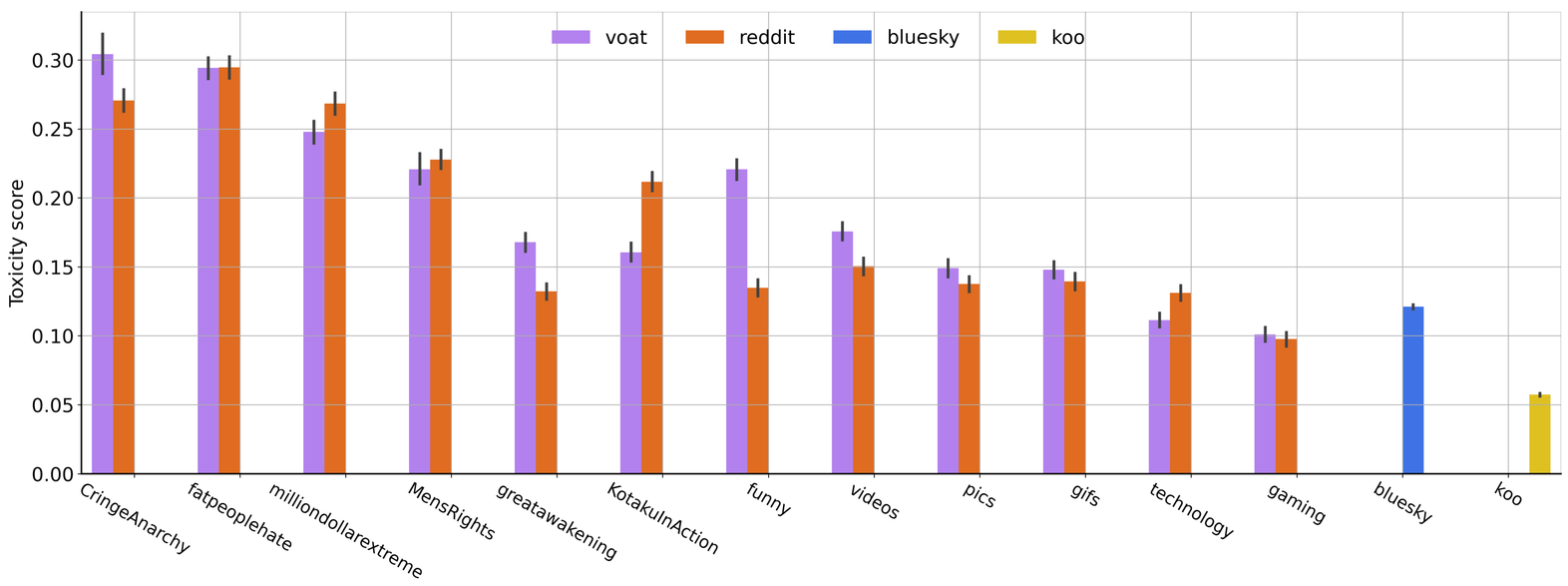
3. Applications

- Cross-platform user behavior analysis and engagement patterns
- Community dynamics during content moderation and policy changes
- Content moderation effectiveness across different platforms
- Toxic behavior propagation and user migration patterns
- Comparative discourse analysis across platform architectures

Dataset statistics

	Time span (years)	Total interactions	Unique users
Bluesky	23 24	2.8 M	0.2 M
Koo	20 23	4.3 M	0.2 M
Reddit	14 17 20	247.6 M	22.6 M
Voat	13 17 20	1.2 M	0.1 M





References

[1] Mekacher, A. and Papasavva, A., 2022, May. " I Can't Keep It Up." A Dataset from the Defunct Voat. co News Aggregator. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 16, pp. 1302-1311).
[2] Mekacher, A., Falkenberg, M. and Baronchelli, A., 2024, May. The koo dataset: An indian microblogging platform with global ambitions. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 18, pp. 1991-2002).
[3] Failla, A. and Rossetti, G., 2024. "I'm in the Bluesky Tonight": Insights from a year worth of social data. PloS one, 19(11), p.e0310330.

Acknowledgment

This research was supported by the Science Fund of the Republic of Serbia, 7416, Topology-derived methods for the analysis of collective trust dynamics - CTRUST.

Data processing was performed on the PARADOX-IV supercomputing facility at the Scientific Computing Laboratory, National Center of Excellence for the Study of Complex Systems, Institute of Physics Belgrade.