Extended Abstract: New Narrative Annotation Infrastructure in the Modern Social Media Ecosystem

Anonymous authors
Paper under double-blind review

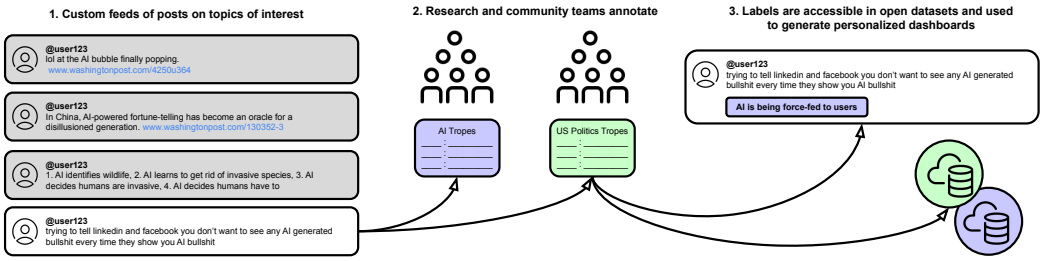

Figure 1: System overview of label contributors and their effect on the feed. Independent teams of annotators (including researchers, community members, and other volunteers) can flag posts for processing and then annotate posts using label sets of narrative tropes. Posts can be flagged anywhere on Bluesky and/or through a custom, topic-focused feed. These labels are then propagated to both personalized dashboards and to open datasets for other researchers to use and build upon.

The modern social media ecosystem is in crisis. Recent events have created a stranglehold over most social media, with the purchase of Twitter by Elon Musk, TikTok's internal censorship and threatened U.S. ban, moderation changes at Meta, and the closing of research APIs supporting an online information ecosystem where misinformation and disinformation can thrive. Attempts to check the spread of inaccurate content have not met with much success. Technical challenges, such as the automatic detection of misinformation at scale, make tackling the problem difficult for engineers; expert fact checkers are not always trusted and have limited capacity to moderate entire social media platforms, let alone the whole of the internet; and crowdsourced projects like Community Notes (formerly Birdwatch), while promising, are gameable and limited in scope and speed (Wojcik et al., 2022; Allen et al., 2022; Borenstein et al., 2025).

But more importantly, battling to correct individual pieces of misinformation is perhaps always a losing game, one that is simply not possible to win; not just because of the scale of the problem but because the larger persuasive goal of this misinformation might have nothing to do with the *information* itself but instead its contribution to (1) a larger weakening of trust in all information and (2) the construction of shared, potentially harmful *narratives* that are harder to pin down and combat. In the words of Alice Marwick, director of the nonprofit research institute Data & Society, "*The problem is less about 'units of facts'... The problem is with these big, sticky stories*" (Clarke, 2025).

Studying such stories, whose specifics might vary but whose shapes can spread quickly and stubbornly across communities, will need the combined expertise of researchers from many different disciplines. However, there is little shared infrastructure to facilitate ongoing joint ownership and collaboration between all narrative stakeholders; narrative labels, such as media frames (Card et al., 2015), moral foundations (Graham et al., 2013), and tropes (Wright et al., 2024), exist in separate frameworks far away from the journalists, community members, engineers, and other researchers who might be able to contribute to such labeling projects. We need large, shared experiments and resources that enable new collaborations.

In this extended abstract, we propose an ambitious research plan to begin to address some of these challenges. This experiment will focus on a single social media platform — Bluesky — which uses an open protocol and supports researcher- and user-driven projects. We outline

here a set of requirements and design plans to create a **collaborative narrative labeling pipeline** on Bluesky. This pipeline will bring together users and researchers to label the kinds of "big, sticky stories" that spread online by collaboratively annotating Bluesky posts with narrative labels.

Narratives and the health of information ecosystems Addressing misinformation in online spaces has traditionally been carried out by human agencies accompanied by automated fact-checking and claim verification systems to determine the veracity of individual posts Guo et al. (2022); Gupta et al. (2021); Warren et al. (2025). But another approach to misinformation in online communities has been to consider shared, high-level narratives as vectors, rather than specific statements whose veracity about specific people and events can be fact-checked. The field of *computational folkloristics* (Abello et al., 2012) has used this approach to study emerging conspiracy theories and their spread across the internet (Shahsavari et al., 2020; Tangherlini et al., 2020). These studies detect "underlying narrative frameworks" and the connections between them (Shahsavari et al., 2020). Similarly, *sensemaking* research, which examines how people construct shared knowledge, has included studies of when this sensemaking process goes awry; in these cases, "facts" are not always the issue but rather rumors and conspiracy theories that spread higher level "misinterpretations and mischaracterizations" (Starbird, 2024). Having access to high quality frames and being able to apply frames appropriately could be the difference between experts and those more susceptible to misinformation; and disinformation can be seen as manipulation of the sensemaking (and frame application) process (Klein et al., 2007; Starbird, 2024).

Prototype design Bluesky[1] is a new social media platform built using the AT Protocol,[2] a decentralized protocol for large-scale social web applications. Bluesky's popularity and its open and "hackable" structure make it ideal for computational social experiments, such as the one we propose here. Bluesky labels[3] can be designed and applied by any user who chooses to run a *moderation service*. Other users may then subscribe to any of these services, which can then hide, warn, or simply display a label for relevant content; the labels are all opt-in. For example, one such service adds labels to the profiles of U.S. politicians with the names of their largest corporate donors.

Our working prototype uses the open-source Ozone[4] library, built on top of skyware and ATProto, to provide a usable interface. Figure 1 shows an overview of our pipeline:

1. Researchers, community members, journalists and others can propose their own topic-specific trope annotation tasks through a formal process. Once approved, these will form annotation teams.
2. Annotators (researchers, community members, general users) who subscribe to the moderation service can flag posts and provide free text rationales and labels.
3. Flagged posts enter a moderation queue for trusted members of the annotation team, who can iteratively cluster and assign labels.
4. Assigned labels and the associated Bluesky post IDs will be immediately available to researchers, builders, and general users via downloadable datasets and personalized dashboards. Data will need to be "re-hydrated" each time it is used, to preserve users' agency to edit and delete their data.

In designing this pipeline, we envision an open infrastructure that will allow for collaboration across narrative research groups and a sharing of annotation resources. All work will be done openly, with both resources and results shared with the community. This project will also include social media users, inviting them into the annotation process and allowing them to own their annotations and data while also sharing with researchers and other users and builders. While the scope of this project is ambitious, it is urgent that more open, transparent, and centralized tools are designed for the new social media ecosystem.

[1] https://bsky.app/

[2] https://atproto.com/

[3] https://docs.bsky.app/docs/advanced-guides/moderation

[4] https://github.com/bluesky-social/ozone

References

James Abello, Peter Broadwell, and Timothy R. Tangherlini. Computational folkloristics. July 2012. URL https://doi.org/10.1145/2209249.2209267.

Jennifer Allen, Cameron Martel, and David G Rand. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in twitter's birdwatch crowdsourced fact-checking program. In *Proceedings of CHI 2022*, New York, NY, USA, 2022. URL https://doi.org/10.1145/3491102.3502040.

Nadav Borenstein, Greta Warren, Desmond Elliott, and Isabelle Augenstein. Can Community Notes Replace Professional Fact-Checkers? *Arxiv*, 2025. URL https://arxiv.org/abs/2502.14132.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. The Media Frames Corpus: Annotations of Frames Across Issues. ACL, 2015. URL https://doi.org/10.3115/v1/p15-2072.

Laurie Clarke. 'Nobody was tricked into voting for Trump': Why the disinformation panic is over. *Politico*, January 2025. URL https://www.politico.eu/article/nobody-tricked-vote-donald-trump-disinformation-panic-over/.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in experimental social psychology*. Elsevier, 2013.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A Survey on Automated Fact-Checking. *TACL*, 2022. URL https://aclanthology.org/2022.tacl-1.11/.

Shreya Gupta, Parantak Singh, Megha Sundriyal, Md. Shad Akhtar, and Tanmoy Chakraborty. LESA: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. In *EACL*, April 2021. URL https://aclanthology.org/2021.eacl-main.277/.

Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. A data–frame theory of sensemaking. In *Expertise out of context*. Psychology Press, 2007.

Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R Tangherlini, and Vwani Roychowdhury. Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of computational social science*, 2020.

K Starbird. Facts, frames, and (mis)interpretations: Understanding rumors as collective sensemaking, 2024. URL https://www.cip.uw.edu/2023/12/06/rumors-collective-sensemaking-kate-starbird/.

Timothy R Tangherlini, Shadi Shahsavari, Behnam Shahbazi, Ehsan Ebrahimzadeh, and Vwani Roychowdhury. An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, pizzagate and storytelling on the web. *PloS one*, 2020.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. Show Me the Work: Fact-Checkers' Requirements for Explainable Automated Fact-Checking. In *ACM Conference on Human Factors in Computing Systems (CHI 2025), to appear*, 2025. URL https://arxiv.org/abs/2502.09083.

Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M. B. Fallin Hunzaker, Keith Coleman, and Jay Baxter. Birdwatch: Crowd Wisdom and Bridging Algorithms Can Inform Understanding and Reduce the Spread of Misinformation, 2022.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge J. Belongie, and Isabelle Augenstein. LLM Tropes: Revealing Fine-Grained Values and Opinions in Large Language Models. In *EMNLP Findings*, 2024. URL https://aclanthology.org/2024.findings-emnlp.995.