

Look What You Made Us Cluster: Hate Narrative Extraction from Reddit Discourse

Annabelle K. L. Chua¹, Forster J. Khoo¹, Joel C. R. Tan¹, Huey Ting Ang¹,
Kheng Hwee Tan¹, Joel Y. A. Sim², Shirley W. H. Ow², Ria Mundhra², Elsie
C. K. Toh², Youfeng Xu, and Lynnette H. X. Ng²

¹ DSO National Laboratories, 14 Science Park 118226, Singapore

² Defence Science and Technology Agency, 1 Depot Road 109679, Singapore

Abstract. Narrative extraction allows us to identify online hate narratives, supporting the construction of rigorous detection systems. Existing computational approaches, however, are limited in precision as they rely on semantic representations, which tend to capture only surface-level meaning. To detect more precise and interpretable narratives, we present an extraction pipeline that represents narratives as entity-evaluation pairs. Narratives are extracted using a Large Language Model (LLM) reasoning process that extends Aspect-Based Sentiment Analysis, identifying the aspect, classifying its judgement type as the basis for evaluation, and deriving the evaluation accordingly. Extracted narratives are then clustered using Leiden, following which clusters are resolved to an intended level of granularity through an LLM-guided refinement process. We illustrate this narrative pipeline with English Reddit comments from 2024 that criticize Taylor Swift, analysing a representative cluster that exhibits hate speech patterns to demonstrate its interpretive value.

Keywords: narrative extraction · aspect-based sentiment analysis · community detection · graph clustering · large language models

1 Introduction

Narrative detection can serve as an important tool for understanding how online discourse spreads and shapes attitudes over time, with particular relevance to online hate, which can emerge from sustained narratives that accumulate gradually across seemingly harmless posts [2]. For instance, narrative similarity has helped identify coordinated user networks, with the disseminated narratives examined to distinguish the agendas of different user groups [18]. More broadly, this illustrates the potential of narrative detection to reveal patterns of harmful online activity, since hate narratives can similarly accumulate and spread through networks of users. However, many existing detection methods built on semantic similarity tend to capture only surface-level patterns and may struggle to identify narratives that rely on indirect or implicit language. We therefore propose a framework that uses LLMs to extract and aggregate online narratives. This framework can help capture how narratives shape audience perceptions of social entities, supporting more effective platform governance.

Narratives are characterizations of individuals, groups, or organizations. When clustered into broader patterns, narratives form meta-narratives: overarching storylines that reflect underlying assumptions and beliefs [3, 16, 23]. Building on this, we introduce our narrative extraction pipeline, which identifies entity-specific narratives by combining aspect-based sentiment analysis with judgement frame analysis. These narratives are then clustered using graph-based community detection and structured LLM reasoning to surface a higher-level landscape of meta-narratives. We demonstrate this pipeline on celebrity discourse, given its high volume and the associated risks of online bullying and reputational harm. Specifically, we apply it to English Reddit comments from 2024 discussing Taylor Swift. Beyond this application, the pipeline can be adapted to other social categories where online hate detection might be a concern.

2 Background

Entity Extraction and Resolution. Entity extraction and resolution are typically addressed by separate models, from statistical and deep learning models for extraction [13] to similarity-based matching, blocking, and clustering for resolution [4]. Recent advances show that LLMs can perform extraction directly, achieving performance comparable to fully supervised baselines and outperforming them in low-resource and few-shot settings [26]. This motivates our investigation into whether a single LLM can also perform resolution reliably, and do so jointly with extraction.

Narrative Extraction. Existing computational approaches to narrative analysis often focus on identifying what narratives exist, without considering how they are framed, limiting the precision of narrative extraction [1, 25]. Such approaches typically rely on semantic representations that capture surface-level similarity, which can produce outputs that are difficult to interpret [14]. Previous work has attempted to address this by adding additional structure to clusters, such as classifying narratives into predefined story types [17]. However, the reliance on manual annotation limits the scalability of such approaches. By deriving discourse structure from theory rather than inferring it after clustering, we extract narratives more precisely and interpretably.

3 Methodology

3.1 Pipeline Overview

Figure 1 illustrates our three-stage pipeline. Given an input document, we identify entities and resolve them to their canonical forms (Section 3.2). For each resolved entity, we then extract narrative features that characterize how the entity is portrayed within the document (Section 3.3). Finally, these features are grouped into buckets and clustered using community detection. The resulting clusters are then refined with an LLM to produce coherent meta-narratives

(Section 3.4). We applied our pipeline to the “Exorde Social Media One Month 2024” dataset [6], analysing English Reddit comments in the Entertainment category. We focused our analyses on “Taylor Swift”, one of the top extracted entities. String matching on “Taylor” or “Swift” yielded 13,812 candidate records.

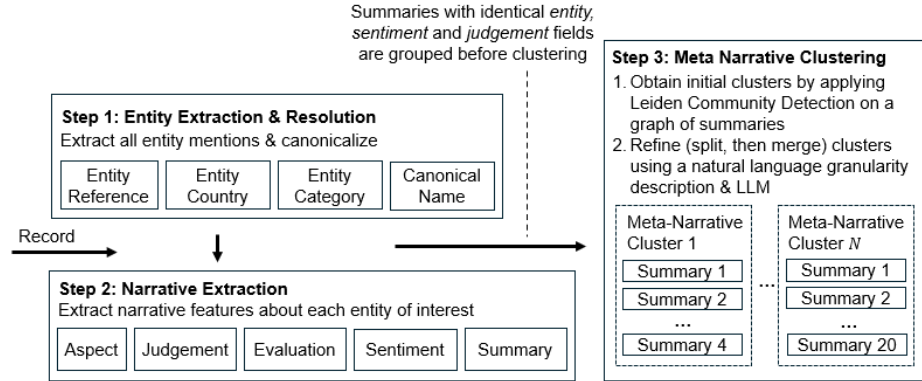


Fig. 1: Narrative Pipeline Logic.

3.2 Entity Recognition and Resolution

Each comment is passed to an LLM, which extracts the list of entities it contains using a single prompt. For each entity, the LLM records its reference (the form in which it appears in the text) and determines its country of residence, along with a category (e.g., individual, group, or organisation), which together support resolving the entity to its canonical form. Resolving entities to a canonical form reduces fragmentation from variations in naming, abbreviations, and indirect references, forming coherent entity groups for downstream narrative analysis. Following this process, the 13,812 candidate records identified in Section 3.1 were filtered down to 6,715.

3.3 Narrative Extraction

Narratives are centered on entities—individuals, groups, and organizations—whose actions, attributes, or experiences form the subject of the story. Following Page’s [20] approach to identifying narratives in short online content, each narrative is defined by two key components: an *entity* and an *evaluation* expressed towards that entity. Evaluations capture attitudes conveyed through praise, criticism, or condemnation [11, 15, 19], shaping audience perception of the entities being discussed [20, 22].

Evaluations are further characterized by their *judgement type* [10, 11], which is the basis on which the evaluation is made. We identify ethics (sincerity, honesty,

or morality) and capability (skill or dependability) as two dominant judgement types for publicly known entities, aligning with the warmth and competence dimensions of social perception identified by the Stereotype Content Model [5, 7]. We also include appearance as a third judgement type, as celebrity and public-figure discourse is heavily shaped by mass-mediated imagery and visual appeal, and appearance-based online hate remains a prevalent yet relatively understudied harm [8]. A breakdown of judgement type definitions provided to the LLM in the narrative extraction prompt is presented in Table 1.

Table 1: Judgement Type Definitions in Our Narrative Extraction Prompt

Judgement Type	Definition
Ethics	Judgements about morality, integrity, honesty, sincerity, kindness, hypocrisy, humility, abuse of power or trust, or harm to others. Some examples include being genuine, dishonest, having double standards, lacking integrity, insincerity.
Capability	Judgements about competence, skill, intelligence, knowledge, ability to deliver, or producing high quality work. This would include characteristics related to the job of the person, e.g., a musician/singer who produces good quality music.
Appearance	Judgements which are about looks and physical appearance. Examples include being beautiful, sexy, handsome, unattractive.

Narrative Extraction is performed by prompting an LLM (Qwen3-30B-A3B-Instruct-2507-FP8) with a document-entity pair (entity obtained from Section 3.2) and a small set of few-shot examples. In a single generation, the model performs a multi-step reasoning procedure that extends Aspect-Based Sentiment Analysis (ABSA) [12] to uncover the evaluation expressed towards the entity.

In the first step, the model identifies the aspect—defined as a specific attribute or characteristic of the entity discussed (e.g., leadership). Next, it determines the type of judgement (ethics, capability or appearance) being expressed, if any. This step serves as a useful checkpoint: the presence of a judgement signals that an evaluation is likely being made, reducing both false positives from non-judgemental descriptions and false negatives from subtle, implicit evaluations based on a judgement type. Finally, the LLM assesses the overall evaluation made towards the entity, as well as a sentiment score (on a scale of -2 to 2, for more resolution when monitoring narratives; collapsed to discrete labels of positive, neutral, or negative for bucketing before clustering) reflecting its direction. Thereafter, the LLM condenses the entity, aspect, and evaluation into a single summary field, which can be used for downstream clustering and analyses. The features extracted at this stage are summarized in Table 2, and an illustrative example of the pipeline outputs is given below.

Example Output. The following comment serves as an input to the narrative extraction pipeline.

Table 2: Summary of Extracted Outputs in Narrative Detection

Feature	Definition
Aspect	The specific characteristic of the entity that is being discussed
Judgement	The underlying basis on which an evaluation is made
Evaluation	Attitudes towards the entity’s behaviour or actions
Sentiment	Sentiment of evaluation towards the entity
Summary	Summarize the evaluation relating to the entity’s aspect

“Sick of hearing about Taylor Swift and all the lame music coming out.”

The pipeline then produces (1) the aspect, **music quality**; (2) the judgement type, **capability**; (3) the evaluation, **lame music coming out**; (4) a **negative sentiment** toward Taylor Swift; and (5) a summary, **Taylor Swift’s music is criticised as lame and unimpressive**.

3.4 Meta-Narrative Extraction

Semantic clustering approaches such as BERTopic group documents by embedding similarity alone [9]. This is inadequate for narrative clustering, as each cluster must be coherent in both the entity it concerns and the sentiment expressed towards it. We address this by grouping summaries from Section 3.3 into buckets sharing a common entity and sentiment polarity. Summaries within each bucket are embedded and clustered using the Leiden algorithm [24], which provides an efficient and scalable approach to community detection for large-scale datasets. Clusters are then labelled and refined through the semantic labelling and cluster refinement stages described below.

Semantic Labelling For each cluster, up to 100 centroid-nearest texts are sampled, providing sufficient representation while remaining within practical LLM input limits. These are passed to an LLM to produce a cluster title and summary, and the title is embedded to obtain a cluster-level embedding used in downstream refinement.

Cluster Refinement Leiden yields structurally sound clusters, but two issues may persist: cluster titles may be near-identical when the underlying texts describe similar opinions, or conversely, overly general when a cluster comprises texts spanning distinct subtopics. We address both through a refinement stage, where an LLM determines whether a cluster should be split, if its title is overly general, or merged with another, if their titles are overly similar. Both decisions are governed by a user-supplied natural-language granularity description, ensuring that splitting and merging operate at the intended level of detail.

Cluster Splitting. Clusters with at least six texts become candidates for this phase. We pass these candidates’ title, summary, and a sampled subset of their texts to an LLM alongside the granularity description, to determine whether the

cluster is overly general, covering multiple distinct subtopics. Clusters identified as such are sub-clustered using Leiden at a resolution of 2.0, above the default resolution, to obtain finer-grained clusters which are then relabelled.

Cluster Merging. We construct a cluster similarity graph where nodes are clusters and edges connect pairs with cosine similarity ≥ 0.75 , computed over cluster-level embeddings, comparing each cluster against its 3 nearest neighbors to limit inference cost. Each candidate edge is passed to the LLM with the granularity description, which determines whether the cluster pair is sufficiently similar for merging. Edges deemed sufficiently similar are retained and the rest pruned. Connected components of the resulting graph define merge groups, which are then relabelled in a single LLM call. Pruned edges with cosine similarity ≥ 0.80 undergo a *relabel* call that produces disambiguating titles and summaries, or leaves the pair unchanged if labels are already sufficiently distinct.

4 Results and Discussion

4.1 Performance of Entity Extraction and Resolution

A dataset to evaluate the performance of the entity extraction and narrative extraction stages was built by manually coding a total of 90 outputs randomly sampled from the Taylor Swift entity bin, with 30 outputs selected for each judgment type to ensure the reliability of the results. Out of these 90 outputs, mentions of Taylor Swift were accurately identified and assigned to the canonical bin in 89 instances for an accuracy of 98.9%.

4.2 Performance of Narrative Extraction

The 90 sampled outputs in the evaluation dataset from Section 4.1 were then annotated by three human raters, all of whom were trained using a common coding scheme. Two independent human raters coded the dataset separately. A third rater then resolved disagreements between the two coders to establish a human ground-truth dataset against which the LLM’s ratings were compared. The validation was conducted sequentially to mirror the dependency structure of the pipeline, where each feature was validated only when its predecessor was correctly identified (Figure 1). For example, sentiment was only assessed within correctly identified evaluations. This ensures that at each stage the LLM and the human rater are assessing the same narrative, and feature-level metrics should be interpreted accordingly as conditioned on correct upstream identification.

Table 3: Results for Entity, Aspect and Evaluation.

Entity (%)	Aspect within accurate entities (%)	Evaluation within accurate aspect and judgement (%)
98.9	75.0	92.5

Subjective categorical variables (e.g., evaluation, aspect) were evaluated differently from fixed-category variables (e.g., judgement, sentiment). For the former, we treat disagreements between coders as natural variation in human interpretation rather than error [21], instead assessing whether the LLM’s output constituted a defensible, textually supported interpretation. Where disagreement remained, a third coder made the final call. Fixed-category variables, by contrast, were independently coded by annotators and compared directly against the LLM’s outputs.

Within correctly identified entities ($N = 89$), aspect had an accuracy of 75% (Table 3). Within accurately identified aspects and judgements ($N = 65$), evaluation accuracy was 92.5%. It should be noted that feature-level metrics are computed on a shrinking denominator at each stage, which presents an optimistic picture of accuracy in isolation. To provide a holistic estimate of overall performance, we also report end-to-end accuracy by counting narratives where all features are correctly identified: of 90 narratives extracted across our sample of entities, 68.9% were fully accurately extracted.

For judgement type, the LLM demonstrated good classification performance, achieving a macro-averaged F1 score of 1.0 across the three judgement type categories. Interrater reliability for sentiment ratings was assessed with a two-way absolute agreement, single-measure intraclass correlation coefficient. Results indicated excellent reliability between the LLM and ground truth, $ICC = .93$, 95% CI [.89, .96], $p < .001$.

While the pipeline demonstrates reasonable performance, the presence of erroneous features suggests that conclusions drawn from individual narrative extractions should be interpreted cautiously. We anticipate that genuinely recurring narratives will consolidate into meaningful meta-narratives, while erroneous narratives, being idiosyncratic, are unlikely to cluster consistently enough to form a meta-narrative. The metrics reported here should therefore be interpreted as a preliminary performance benchmark rather than a definitive constraint on the pipeline’s utility.

4.3 Performance of Meta-Narrative Extraction

Figure 2 presents a meta-narrative recovered by the pipeline. The example illustrates the pipeline’s ability to recover semantically coherent opinion clusters. Although the constituent comments draw on distinct points of comparison—her “girly” image, her body shape, and comparisons to other artists—they converge on a shared narrative that Taylor lacks sex appeal despite being objectively attractive. This offers an appropriately granular view of how she is perceived, neither fragmenting into narrow sub-areas such as commentary on a single physical feature, nor collapsing into broad, undifferentiated criticism of her appearance or character. Such thematic specificity could be lost in a broader semantic cluster spanning multiple entities or sentiment polarities. Individual comments in this cluster often read as superficially positive or neutral, prefacing their critique with a compliment, as in “I absolutely love Taylor Swift...” or “Taylor Swift is objectively pretty...”. Evaluated on their surface sentiment alone, such comments

might not be flagged by post-level hate detection techniques that assess comments in isolation. It is only once our pipeline extracts the underlying evaluation embedded within each comment, and aggregates it across the cluster, that the broader pattern of characterization becomes visible, along with its potential to shape perception and cause reputational harm over time.

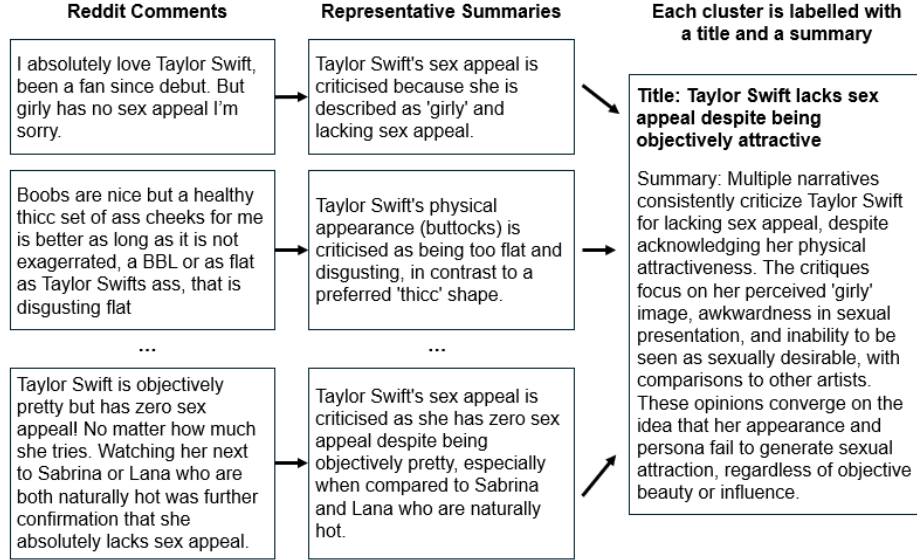


Fig. 2: Sample Meta-Narrative Illustration.

5 Conclusions and Future Work

This paper presented a pipeline for narrative extraction and clustering that facilitates more precise and interpretable narratives. This then provides a useful avenue for online narrative monitoring, supporting efforts to detect and understand hate narratives as they emerge.

Several limitations suggest directions for future work. First, clustering quality is bounded by the accuracy of upstream entity and sentiment extraction. Errors at this stage propagate into bucket composition and cannot be recovered downstream. Second, the labelling and refinement stages are coupled to the behaviour of the underlying LLM. Both model sensitivity and the non-determinism inherent in LLM calls introduce variation in split and merge decisions, and future work should explore guardrails such as confidence thresholds or ensemble voting across runs to improve consistency. Third, the pipeline processes a static corpus, and extending it to support incremental clustering over evolving document streams would be a natural step toward studying long-term opinion dynamics

and coordinated narrative activity in social discourse. Finally, there is a need to fully operationalise the detection of hate narratives by defining criteria at the cluster level. Possible methods include identifying clusters with the lowest average sentiment, or identifying cluster labels with concerning semantic indicators.

References

1. Alieva, I., Kloos, I., Carley, K.M.: Analyzing Russia’s propaganda tactics on Twitter using mixed methods network analysis and natural language processing: A case study of the 2022 invasion of Ukraine. *EPJ Data Science* **13**(1), 42 (2024). <https://doi.org/10.1140/epjds/s13688-024-00479-w>
2. Almagro, M., Vieites, C., Chakkour, H., Sancho, A.: How can hate narratives be tracked in online environments? a corpus-based study. *Frontiers in Communication* **11**, 1743196 (2026). <https://doi.org/10.3389/fcomm.2026.1743196>
3. Ang, S.: I am more Chinese than you: Online narratives of locals and migrants in Singapore. *Cultural Studies Review* **23**(1), 102–117 (2017). <https://doi.org/10.5130/csr.v23i1.5497>
4. Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., Stefanidis, K.: An overview of end-to-end entity resolution for big data. *ACM Computing Surveys* **53**(6) (2020). <https://doi.org/10.1145/3418896>
5. Cuddy, A.J., Fiske, S.T., Glick, P.: Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. In: Zanna, M.P. (ed.) *Advances in Experimental Social Psychology*, vol. 40, pp. 61–149. Elsevier Academic Press (2008). [https://doi.org/10.1016/S0065-2601\(07\)00002-0](https://doi.org/10.1016/S0065-2601(07)00002-0)
6. Exorde Labs: Multi-source, multi-language social media dataset. Exorde Labs (2024), <https://www.exordelabs.com/>
7. Fiske, S.T., Cuddy, A.J.C., Glick, P., Xu, J.: A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology* **82**(6), 878–902 (2002). <https://doi.org/10.1037/0022-3514.82.6.878>
8. Grasso, F., Valse, A., Micheli, M.: Body-shaming detection and classification in Italian social media. In: Rapp, A., Di Caro, L., Mezziane, F., Sugumaran, V. (eds.) *Natural Language Processing and Information Systems: NLDB 2024, Lecture Notes in Computer Science*, vol. 14762. Springer (2024). https://doi.org/10.1007/978-3-031-70239-6_18
9. Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794* (2022), <https://arxiv.org/abs/2203.05794>
10. Hansson, S., Fuoli, M., Page, R.: Strategies of blaming on social media: An experimental study of linguistic framing and retweetability. *Communication Research* **51**(5), 467–495 (2024). <https://doi.org/10.1177/00936502231211363>
11. Hansson, S., Page, R., Fuoli, M.: Discursive strategies of blaming: The language of judgment and political protest online. *Social Media + Society* **8**(4) (2022). <https://doi.org/10.1177/20563051221138753>
12. Hua, Y.C., Denny, P., Wicker, J., Taskova, K.: A systematic review of aspect-based sentiment analysis: domains, methods, and trends. *Artificial Intelligence Review* **57**, 296 (2024). <https://doi.org/10.1007/s10462-024-10906-z>
13. Keraghel, I., Morbieu, S., Nadif, M.: Recent advances in named entity recognition: A comprehensive survey and comparative study. *arXiv preprint arXiv:2401.10825* (2024), <https://arxiv.org/abs/2401.10825>

14. Ma, Y., Xiao, C., Yuan, C., Veer, S.N.V.D., Hassan, L., Lin, C., Nenadic, G.: CAST: Corpus-aware self-similarity enhanced topic modelling. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 7548–7561. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). <https://doi.org/10.18653/v1/2025.naacl-long.386>, <https://aclanthology.org/2025.naacl-long.386/>
15. Martin, J.R., White, P.R.R.: The language of evaluation. Palgrave Macmillan UK (2005)
16. Moss, S.M., Sandbakken, E.M.: Everybody needs to do their part, so we can get this under control: Reactions to the Norwegian government meta-narratives on COVID-19 measures. *Political Psychology* **42**(5), 881–898 (2021). <https://doi.org/10.1111/pops.12727>
17. Ng, L.H.X., Carley, K.M.: “The coronavirus is a bioweapon”: classifying coronavirus stories on fact-checking sites. *Computational and Mathematical Organization Theory* **27**(2), 179–194 (2021). <https://doi.org/10.1007/s10588-021-09329-w>
18. Ng, L.H.X., Carley, K.M.: Do you hear the people sing? comparison of synchronized url and narrative themes in 2020 and 2023 french protests. *Frontiers in Big Data* **6**, 1221744 (2023). <https://doi.org/10.3389/fdata.2023.1221744>
19. Oteiza, T.: The appraisal framework and discourse analysis. In: Bartlett, T., O’Grady, G. (eds.) *The Routledge Handbook of Systemic Functional Linguistics*, pp. 481–496. Routledge, 1 edn. (2017)
20. Page, R.: Re-examining narrativity: small stories in status updates. *Text & Talk* **30**(4), 423–444 (2010). <https://doi.org/10.1515/TEXT.2010.021>
21. Pavlick, E., Kwiatkowski, T.: Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics* **7**, 677–694 (2019). https://doi.org/10.1162/tacl_a_00293
22. de Saint Laurent, C., Glăveanu, V.P., Literat, I.: Internet memes as partial stories: Identifying political narratives in coronavirus memes. *Social Media + Society* **7**(1) (2021). <https://doi.org/10.1177/2056305121988932>
23. Tannen, D.: We’ve never been close, we’re very different: Three narrative types in sister discourse. *Narrative Inquiry* **18**(2), 206–229 (2008). <https://doi.org/10.1075/ni.18.1.03tan>
24. Traag, V.A., Waltman, L., Van Eck, N.J.: From louvain to leiden: guaranteeing well-connected communities. *Scientific reports* **9**(1), 1–12 (2019). <https://doi.org/10.1038/s41598-019-41695-z>
25. Trimmingham, C., Agarwal, N.: Leveraging topic modeling and toxicity analysis to understand China-Uyghur conflicts. In: Proceedings of the 15th International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2022). Springer, Pittsburgh, Pennsylvania, USA (September 2022), http://sbp-brims.org/2022/papers/working-papers/2022_SBP-BRiMS_Final_Paper_PDF_9417.pdf
26. Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., Wang, G.: Gpt-ner: Named entity recognition via large language models. arXiv preprint arXiv:2304.10428 (2023), <https://arxiv.org/abs/2304.10428>