Mapping Strategic Narratives in Chinese State Media: A Sentence-BERT Analysis of Russia-Ukraine and Israel-Palestine War Coverage

Extended abstracts

1 Introduction and Contribution

This research-in-progress (RIP) studies how Chinese state media computationally organize strategic narratives about two high-salience wars: Russia-Ukraine and Israel-Palestine. Strategic narrative theory argues that political actors do not merely describe events; they assign roles, causes, victims, and legitimate solutions within broader stories about order and responsibility [1]. For social computing, the methodological challenge is to measure such narrative organization at scale without reducing it to keyword counts or treating embeddings as uninterpretable black boxes [2].

The project asks: how do Chinese state media adapt strategic narratives across conflicts that differ in geopolitical stakes, humanitarian salience, and China's diplomatic room for maneuver? The current submission reports a pilot analysis of 6,497 CGTN articles to validate the measurement design. The full RIP will expand the corpus beyond CGTN to China Daily, People's Daily/People's Daily Online, and Xinhua News Agency, with an expected final corpus above 10,000 articles. The contribution is a scalable but theory-guided pipeline for comparing state-media narratives across outlets, wars, and time.

2 Data and Computational Design

The pilot corpus contains publicly available CGTN English-language articles collected with conflict-related keywords and screened for relevance. The Russia-Ukraine window runs from February 2022 to January 2023; the Israel-Palestine window runs from October 2023 to September 2024. Cleaning removed captions, source labels, navigation text, wire-service markers, and repeated boilerplate. The next data-collection stage will harmonize metadata across outlets, deduplicate syndications, and balance articles by war, outlet, and month so that narrative differences are not driven only by CGTN's editorial profile or by corpus-size imbalance.

Component	Specification
Current pilot	CGTN only: 4,239 Ukraine articles and 2,258 Israel-Palestine articles.
Planned expansion	CGTN, China Daily, People's Daily/People's Daily Online, and Xinhua News Agency; expected corpus >10,000 articles.
Representation	SBERT embeddings [3] at sentence/paragraph level, aggregated to article-level and kept in a shared semantic space.
Discovery	UMAP + HDBSCAN for inductive themes; c-TF-IDF and nearest texts for interpretation.
Narrative measurement	Seven prototype-centroid dimensions following narrative-similarity logic [4]: victimhood, China's role, UN/international community, U.S./Western responsibility, conflict roots, solution, and long-term impact.
RIP validation	Prototype coding agreement in pilot: 86.4%; alpha = .781; kappa = .776. Next: manual labeled subset, model-variant checks, outlet-balanced estimates, and bootstrap confidence intervals.

Table 1. Computational design of the extended abstract.

3 Pilot Findings

The pilot results are not presented as final evidence, but they show that the measurement pipeline can recover interpretable differences. CGTN Gaza coverage is more event-driven and humanitarian, with themes around ceasefire negotiations, civilian casualties, international coordination, regional escalation, and China's peace mediation. Ukraine coverage is more geopolitically ordered, with themes around daily war progression, sanctions consequences, energy and food insecurity, NATO/U.S. involvement, and diplomatic settlement. This difference matters because both corpora come from the same outlet; the variation is therefore not only topical but narrative.

Prototype similarity scores sharpen this contrast and guide the larger RIP design. In the Israel-Palestine pilot corpus, the strongest mean dimensions are UN/international community role (M = .396), war solution (M = .369), U.S./Western responsibility (M = .319), and humanitarian victimhood (M = .308). In the Russia-Ukraine corpus, the strongest dimensions are long-term war impact (M = .323), war solution (M = .319), and U.S./Western responsibility (M = .318), while humanitarian victimhood is much lower (M = .196). Dominant-dimension distributions follow the same pattern: Gaza paragraphs most often align with UN/international community role (33.7%), humanitarian victimhood (21.1%), and war solution (16.3%); Ukraine paragraphs most often align with long-term impact (25.0%), U.S./Western responsibility (18.6%), and war solution (18.3%).

The temporal analysis suggests different trajectories to be tested in the expanded multi-outlet corpus. In Gaza coverage, the UN/international community dimension remains dominant across the observed period, while China's constructive-role narrative rises in the middle and later months. In Ukraine coverage, the early phase emphasizes U.S./Western responsibility, but from April 2022 onward the center of gravity shifts toward long-term global impact. CGTN thus moves from blame attribution to systemic consequence in Ukraine, while maintaining a more stable humanitarian-multilateral storyline around Gaza.

4 ICSC Relevance and Next Steps

This RIP fits ICSC because it connects large-scale media analytics, social applications of language models, and the societal consequences of computational information systems. It also responds to a common weakness in political text-as-data research: unsupervised clusters are scalable but often hard to interpret, while close reading is interpretable but difficult to scale [2]. The next version will report the expanded multi-outlet corpus, exact embedding checkpoints and hyperparameters, cluster-stability checks, and outlet-by-month comparisons. The final paper will also separate textual organization from claims about editorial intent or audience effects, treating the method as a tool for measuring narrative structure rather than proving causal persuasion. Related work on Chinese state-media strategy [5,6] makes this an important case for testing how computational methods can identify patterned public narratives across international crises.

References

[1] Miskimmon, A., O'Loughlin, B., Roselle, L.: Strategic Narratives: Communication Power and the New World Order. Routledge (2013).

[2] Grimmer, J., Stewart, B.M.: Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis 21(3), 267-297 (2013).

[3] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. EMNLP-IJCNLP, 3982-3992 (2019).

[4] Waight, H. et al.: Quantifying narrative similarity across languages. Sociological Methods & Research 54(3), 933-983 (2025).

[5] Fu, K.W.: Propagandization of policy reversal: China state media's nationalistic narratives of dynamic zeroing and living with the virus. Asian Journal of Communication (2025).

[6] Fan, Y., Pan, J., Sheng, J.: Strategies of Chinese state media on Twitter. Political Communication 41(1), 4-25 (2024).