Supporting Information

Across the Firewall: Foreign Media's Role in Shaping Chinese Social Media Narratives on the Russo-Ukrainian War

Hans W. A. Hanley, Yingdan Lu, and Jennifer Pan

Correspondence to yingdan@northwestern.edu

Contents

S1 Weibo Ukraine War Discourse Dataset

Between February 1, 2022, and April 19, 2022, we used Weibo's advanced search function to collect posts containing the words "Ukraine" (乌克兰) and "Ukrainian" (乌克兰人) in simplified and traditional Chinese, in English, as well as Chinese acronyms for the conflict, e.g., "Ukraine+Russian" (俄乌). We identified 475,231 Weibo posts that met this criteria, made by 160,792 users.

Because not all posts containing the word "Ukraine" actually pertained to the Ukraine War, we manually labeled a random sample of 3,200 posts, with stratified sampling by post-creation date, as related to the Ukraine War or not. Intercoder agreement on the hand-labeled data was high (average pairwise percent agreement = 85.4%; Krippendorff's Alpha = 0.80). This hand-labeled data was then used to fine-tune a pre-trained Chinese BERT model [4], achieving accuracy in the test set of 0.95 and F1 score of 0.97 for classifying posts as war-related or non-war-related. A total of 435,261 Weibo posts from February 1 to April 19, 2022, were identified as related to the Ukraine war, forming what we refer to as the Weibo Ukraine War Discourse dataset.

S2 News Websites Dataset

To identify whether narratives from Chinese, Russian, Ukrainian, and US news ecosystems make their way to Chinese social media, we collected news articles from the most trafficked news websites based in Russia, Ukraine, China, and the US To identify US-based news websites, we used a set of 2,500 US-based news domains labeled in prior work [12, 2, 23]. This list of websites includes sites such as washingtonpost.com, apnews.com, cnn.com, and foxnews.com.

To identify popular and influential Chinese, Russian, and Ukrainian news websites, we combined Amazon Alexa Popularity Ranking data [1] and Common Crawl's Domain rank datasets [15] with website categorization data from Cloudflare.[S1] Namely, we collected the set of most popular websites ranked in Amazon Alexa's top one million websites and Common Crawl's Domain Rank datasets from April 2022, which utilize the top-level domain of each country we were interested in (*i.e.*, .cn, .ua, and .ru).[S2] We then queried Cloudflare's Domain Intelligence API to collect the Cloudflare label for each domain and subsequently gathered domains labeled by Cloudflare as news-related. To facilitate comparison across news ecosystems, we use some of the top 2,500 .cn (Chinese), top 2,500 .ru (Russian), and top 2,500 .ua (Ukrainian) most-trafficked and available news domains in the analysis. To check that the domains we attribute to each country are indeed from that country, we collected the domain-registration data of each of the 10,000 domains to confirm that these websites were registered in the respective countries using

[S1] Specifically, we use the Cloudflare Domain Intelligence API, which provides genre classifications of websites (*i.e.*, not limited to Cloudflare hosted websites [19]).

[S2] We utilized both Common Crawl and Amazon Alexa due to the paucity of Chinese, Ukrainian, and Russian domains in the US-dominated Amazon Alexa list.

`whois` in Python. We release the full list of websites utilized in this work at https://github.com/hanshanley/narrative-influence.

From each news website in our dataset, we collected news articles published between January 1, 2022, and June 1, 2022. To gather this data, we took two main approaches: (1) gathering available web crawls from Common Crawl [22], and (2) extensively crawling each website retrospectively between November 2022 and March 2023. Common Crawl is widely considered the most complete public source of web crawl data. For each website, we downloaded Common Crawl indexed pages[S3] from between January 1, 2022, and January 1, 2023 (CC-MAIN-2023-06, CC-MAIN-2022-49, CC-MAIN-2022-40 CC-MAIN-2022-33, CC-MAIN-2022-27, CC-MAIN-2022-21, CC-MAIN-2022-05), identified the publication date using the Python `htmldate` library, and included HTML pages published on their websites between January 1, 2022, and June 1, 2022. To further expand the Common Crawl dataset, we performed a breadth-first crawl (15 hops from the homepage) of each website to gather the set of HTML pages that are missing from Common Crawl. We performed this scraping procedure by using the Golang library `Colly`.[S4] Once all HMTLs are collected, we then parsed each HTML page to collect the published articles by using the `newspaper3k` library. In several instances (e.g., sputniknews.com) where this approach failed, we built custom parsers based on page-specific HTML elements. This generated 24,136,997 articles, including 4,413,358 articles from 2,500 US news domains, 4,528,525 articles from 2,500 Chinese news domains, 10,080,583 articles from 2,500 Russian news domains, and 5,114,531 articles from 2,500 Ukrainian news domains.

Many of the new websites we collected articles from have articles in languages other than English, Chinese, Ukrainian, and Russian (e.g., CNN has articles in Spanish, the *New York Times* has a dedicated Chinese version, and many Ukrainian websites have a Ukrainian and a Russian version). For our analysis, we only included news articles in the primary language of the website (English for US sites, Russian for Russia-based sites, Ukrainian for Ukraine-based sites, and Chinese for China-based sites). This minimizes double counting, as news domains publishing articles in multiple languages typically post the same content across languages. For example, we would only consider English-language articles from CNN because English is its primary language. Similarly, we would only consider English-language articles from the English-language *Sputnik News* website, which, despite being Russian-operated, targets an English-speaking audience with articles in English. To detect the primary language of each website, we utilized the Python `langdetect` library [21]. After deduplication and the removal of articles that were not in the primary language of a given website, the dataset used for analysis contains 4,011,722 articles from 2,500 US news websites, 1,212,795 articles from 2,500 Chinese news websites, 1,688,256 articles from 2,500 Russian news websites, and 1,342,784 articles from 2,500 Ukrainian news websites.

[S3] https://github.com/lxucs/commoncrawl-warc-retrieval
[S4] https://github.com/gocolly/colly

S3 Preprocessing Weibo Messages and News Articles

For each news article and Weibo message, we removed all URLs, emojis, and HTML tags. After this preprocessing, we segmented each article and Weibo post into its constituent paragraphs. In line with prior work [16], we further separated these paragraphs into smaller passages of up to at most 100 Chinese characters or 100 English, Ukrainian, and Russian words, a medium length that captures context when obtaining an embedding (vector representation) for the narrative present within the passage [11, 10, 16]. We refer to these 100-character Chinese messages and 100-word English, Russian, and Ukrainian messages as *passages*.

S4 Clustering and Interpreting Weibo Narratives

We first embedded (calculated vector representations) each of our Weibo passages into a shared embedding space, so that passages conveying similar content or ideas have high cosine similarity. To do this, we utilized a monolingual version of Mandarin-Chinese BERT [5] trained on semantic similarity tasks [18, 14].[S5] Upon embedding each of our Weibo passages into a shared embedding space, we subsequently clustered these embeddings utilizing DP-Means [6]. DP-Means is a non-parametric extension of the K-means algorithm that does not require the specification of the number of clusters *a priori*. DP-Means, however, requires the setting of a value λ for assigning a given embedding/Weibo post to a given cluster. To determine an appropriate λ, we used a manual human-validation approach. Namely, we outputted different clustering solutions with different λ. For each solution, we randomly sampled 150 narrative clusters for human validation. Native Chinese speakers read all passages in each cluster to determine whether passages referred to a singular, identifiable narrative or to different narratives. Upon finishing this evaluation, we selected a $\lambda = 0.10$ (minimal cosine similarity = 0.90), where over 70% of the 150 clusters contained texts pointing to a singular narrative easily identifiable by humans.

Finally, we performed multi-document summarization utilizing an instruction fine-tuned version of `llama3` [7] with the following prompt: *You work for a news researcher and your job is to create an English summary of Chinese texts. Write a single concise 10-word abstractive summary of the texts, where individual texts are separated by |||||. Focus on the text related to Russia and Ukraine.* This generates human interpretable summaries of each Weibo narrative cluster. The 100 top Weibo narratives, in terms of the number of posts in the narrative, are shown in Figure S1. We release the full list of Weibo narratives at https://github.com/hanshanley/narrative-influence.

[S5] https://huggingface.co/shibing624/text2vec-base-chinese-paraphrase.

Table S1: Auto-Generated Summaries of Weibo Narratives with the Largest Number of Weibo Posts.

No.	Posts	Auto-Generated Summary
1	8129	US manipulates Ukraine into war with Russia for its own interests and benefits.
2	2576	Russian-Ukrainian conflict becomes a protracted war with uncertain long-term outcome.
3	1779	US denies allegations of biological warfare in Ukraine, despite evidence contradicting claims.
4	1681	Ukraine and Russia negotiations are unlikely to yield results due to fundamental disagreements.
5	1366	China maintains a neutral stance on the Ukraine-Russia conflict, with some citizens supporting Russia and others Ukraine.
6	1031	Russian President Putin escalates military action in Ukraine, testing Western resolve.
7	975	The US is manipulating Ukraine to create chaos in Europe and distract from its rivalry with China.
8	884	Ukraine's pursuit of NATO membership sparks conflict with Russia, prompting concerns about regional stability.
9	862	Weibo may display users' real-time locations in comments about Ukraine or Russia.
10	833	Human traffickers target Ukrainian female refugees in Germany, sparking outrage.
11	813	Ukraine's fate is largely self-inflicted, a pawn in global geopolitics.
12	788	US media suggests China could face sanctions if Russia invades Ukraine.
13	771	Russia-Ukraine conflict escalates as NATO's aggressive expansion sparks war.
14	714	Thousands of Ukrainian surrogate babies are being kept in underground facilities.
15	653	Chinese netizens are criticized for making insensitive comments about Ukraine, Ukraine's conflict with Russia.
16	638	US and China tensions escalate as Russia's Ukraine invasion shifts global dynamics.
17	605	China provides food aid to Ukraine, while the US supplies lethal weapons.
18	604	Russian official suggests cancelling Ukraine's foreign debt altogether.
19	588	South Korea announces military aid to support Ukraine's conflict with Russia.
20	585	China urges restraint and diplomacy to resolve Ukraine crisis peacefully.
21	582	Russian-Ukraine conflict fueled by NATO's eastward expansion and US dominance.
22	579	US President Biden is confident Putin plans to attack Ukraine.
23	573	Despite initial setbacks, Russia's military strategy has shifted to focus on Ukraine's eastern region.
24	562	Russia and Ukraine tensions escalate as US and Russia exchange warnings.
25	534	Allegations of US conducting biological experiments on Ukrainian mental patients surface.
26	526	Russia-Ukraine conflict is a US-led strategy to pressure China over Taiwan.
27	517	Russia claims to have destroyed 74 Ukrainian military targets on the ground.
28	510	US and EU face economic instability due to Ukraine conflict escalation.
29	503	US acknowledges Ukraine's existence of biological research facilities amidst controversy.
30	502	15 US biological labs found in Ukraine spark widespread concern.
31	499	Russian and Ukrainian tensions escalate, raising fears of a global conflict.
32	470	China's netizens largely condemn Russia's invasion of Ukraine, urging peace.
33	463	Ukrainian President accuses Western countries of abandoning Ukraine.
34	462	Russia-Ukraine conflict escalates, Ukraine's neutrality key to peace outcome.

No.	Posts	Auto-Generated Summary
35	456	Critics accuse the US of sinister motives in Ukraine's labs.
36	432	Chechen forces are reportedly heading to Ukraine for an unknown mission.
37	413	US President Biden warns of potential Russian invasion in Ukraine soon.
38	404	Most of Ukraine's top 100 richest individuals have fled the country.
39	403	Ukraine rejects Swedish and Austrian neutrality models in the conflict.
40	371	Russia's strategic deterrence enters a heightened state, increasing the risk of nuclear war.
41	357	A Korean social media influencer was denied entry into Poland from Ukraine.
42	357	Ukraine officials report fire at Zaporizhzhia nuclear plant, no radioactive leaks.
43	356	Ukraine's EU membership bid faces challenges due to Russia tensions.
44	354	North Korea attributes Ukraine crisis to US imperialism.
45	353	US repeatedly fuels Russia-Ukraine tensions with unsubstantiated claims.
46	350	Russian forces advance in Ukraine, capturing key areas and equipment.
47	349	Russia launches intense attacks on Ukraine, straining international relations.
48	346	Russian and Ukrainian foreign ministers hold talks in Turkey for the first time.
49	343	Russian President Putin claims Ukraine was established from Russian territory.
50	342	Russian President Putin threatens war if any country sets up a no-fly zone over Ukraine.
51	342	Poland's plan to donate Soviet-made MiG-29s to Ukraine is met with US opposition.
52	339	Ukraine's Chernobyl nuclear plant has completely lost power amid tensions.
53	322	Russia's Ukraine invasion has become a humanitarian crisis situation.
54	320	Ukraine and Russia to hold new peace talks amid ongoing conflict.
55	301	US accepts only 7 Ukrainian refugees amid ongoing conflict.
56	299	Ukraine's president, a former actor, faces criticism for handling the Russia-Ukraine conflict poorly.
57	291	Ukraine's plea for help from South Korea is met with amusement.
58	282	US President Biden mistakenly refers to Ukraine as China.
59	275	Russia and Ukraine's conflict escalates as civilians are drawn into the fighting.
60	271	Ukrainian President Zelensky confirms Ukraine will not join NATO anytime soon.
61	269	Ukraine accepts donations in 4 currencies for military aid.
62	264	Ukrainian President Zelensky's leadership and motives under scrutiny amidst Russia conflict.
63	264	Russia's invasion of Ukraine is a unjust war with Ukraine fighting bravely.
64	257	Russian-Ukrainian relations are being tested and discussed on social media platforms.
65	251	Netizens mock Ukraine's request to join the European Union immediately.
66	250	China begins evacuating its citizens from Ukraine in a safe manner.
67	246	Ukraine dam destruction reports by Chinese satellites lack credibility.
68	244	Russia and Ukraine negotiations show significant progress towards a comprehensive agreement.
69	244	Ukrainian and Russian athletes show unity and wish for global peace.
70	239	US President Biden reaffirms no US military involvement in Ukraine conflict with Russia.
71	233	Russia-Ukraine conflict risks global food crisis and soaring prices.
72	229	Russian and Ukraine hold peace talks in Belarus with uncertain outcomes.

No.	Posts	Auto-Generated Summary
73	228	Ukraine seeks 500 billion USD in financial aid from the G7 alliance.
74	227	Russian military claims to have found evidence of US-funded biolabs in Ukraine researching COVID-19 and bird-borne diseases.
75	223	Ukrainian president Zelensky releases former prisoners to fight in the war.
76	222	Ukraine struggles to resist Russian military advances and diplomatic pressure.
77	215	China urges caution and reflection amid Ukraine tensions with Russia.
78	212	China's internet under attack from foreign networks, primarily from the US.
79	209	Ukraine's official statement also released on other social media platforms.
80	207	Ukrainian President Zelensky confirms presence in Kyiv despite Russian invasion.
81	204	Ukrainian President Zelensky criticizes international inaction on Ukraine's issues.
82	201	Russia-Ukraine app reportedly shows users' geographic locations.
83	200	Russian President Putin denies Ukraine invasion plans, sparking controversy.
84	199	Russia and Ukraine to hold second round of talks on March 2.
85	196	Ukraine begins mobilizing 18-60 year-old reservists for potential war.
86	191	Ukraine's Zelensky seeks alliance against Russia, while Russia expands operations.
87	189	Russian-backed Chechen forces may escalate violence in Ukraine.
88	186	Ukrainian President Zelensky signs law seizing Russian assets in Ukraine.
89	183	Ukrainian forces continue to target Russian military equipment and positions.
90	180	Russia and Ukraine's war is a complex multi-faceted conflict fueled misinformation.
91	180	Russia-Ukraine conflict escalates with ongoing military and diplomatic tensions.
92	178	Ukraine allegedly fires at disputed eastern Ukraine region, sparking controversy.
93	178	Russian President Putin's recognition of Ukraine's eastern regions as independent states escalates tensions.
94	178	Russian and Ukrainian citizens express opposing views on the conflict, with some urging Ukraine to surrender and others supporting Ukraine's resistance.
95	173	US and Ukraine collaborate on biological research, amid Russian concerns and accusations.
96	171	Ukrainian forces launch counterattacks against Russian military positions.
97	170	Ukrainians question the destruction of the world's largest transport aircraft AN225.
98	170	Germany donates 2700 Soviet-made missiles to Ukraine, sparking criticism.
99	170	Ukrainian President Zelensky acknowledges Ukraine cannot join NATO currently.
100	170	Ukraine's EU membership bid gains momentum as EU starts review process.

S5 Multilingual Semantic Search

To identify news article passages that convey similar messages to narratives to our set of Weibo posts, we embedded Weibo and news passages utilizing a multilingual version of MPNet.[S6] Then, we identified the set of Weibo passages that have the highest semantic similarity to the set news article passages. Given the massive volume of our data, we

[S6] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2.

utilize the `FAISS` library [13], a library for performing efficient semantic search, to assign and match news passages to Weibo messages.

While semantic search allows us to identify narratives on Weibo that are semantically similar to multilingual news articles, prior work [8, 20] shows that two passages with high semantic similarity might characterize a particular event in very different ways. For example, two passages with high semantic similarity could be for or against a particular event in the Russo-Ukrainian War. Based on our narrative conceptualization, these two passages should be different narratives. To refine the approach to identifying co-occurring narratives, we train three additional mDeBERTa models to identify *paraphrases* between Mandarin Chinese and Ukrainian, Mandarin Chinese and English, and Mandarin Chinese and Russian. If an article passage paraphrases a specific Weibo passage, we assign the article passage to its corresponding Weibo narrative cluster. In other words, if a Russian news article passage is a paraphrase of a Chinese Weibo passage, we consider it to be part of the same narrative.

To train the three mDeBERTa-based paraphrase models, we draw from the Conference on Machine Translation (WMT-19) dataset of news commentary bitexts in Chinese and English, and in Chinese and Russian.[S7] For each language bitext, we select 25,000 positive paraphrase examples and 25,000 negative paraphrase examples for training. For both validation and testing, we utilize 1,000 positive and 1,000 negative paraphrase examples for each language. Given the absence of Ukrainian from the original WMT-19 dataset, we translate the Russian texts within this dataset utilizing a pre-trained model from Helsinki NLP[S8] to obtain a corresponding Chinese-Ukrainian bitext dataset. This particular model achieves a BLEU score [17] of 64.0 on the Tatoeba test set [24] for translating Russia to Ukrainian, indicating fluent and often above human-quality translations. Across each of our models, we achieve 93.1%, 95.0%, and 90.3% accuracy on the respective test sets for identifying cross-lingual paraphrases in Chinese-English, Chinese-Russia, and Chinese-Ukrainian. For our set of Chinese news article passages, we utilize the previously determined human-validated similarity threshold for considering a given Chinese news passage to have the same narrative as a Weibo passage.[S9]

S6 Pointwise Mutual Information

Pointwise mutual information (PMI) is an information-theoretic measure for identifying the associations among two entities by comparing their joint probability (the probability of entities occurring together) and individual probabilities (probability if entities were independent) [3]. Here, the two entities are Weibo narratives and news ecosystems, and we utilize the metric to determine the proportion of narratives that are most associated with

[S7] https://opus.nlpl.eu/News-Commentary.php

[S8] https://huggingface.co/Helsinki-NLP.

[S9] We release the weights and training code for these models at https://github.com/hanshanley/narrative-influence.

each news ecosystem based on the frequency of articles matching each Weibo narrative in each ecosystem. The PMI of a particular Weibo narrative cluster c_i with regards to a given news ecosystem (*i.e.*, Russian, Ukrainian, US, Chinese) E_j is calculated as follows:

$$PMI(c_i, E_j) = log_2 \frac{P(c_i, E_j)}{P(c_i)P(E_i)}$$

where P is the probability of occurrence. We incorporate a scaling parameter α to the raw counts to prevent single counts in each news ecosystem from having the highest PMI values. We utilize a baseline count of 1 (α =1) [25]. In other words, adding α =1 to each count smooths the probabilities, preventing rare counts from disproportionately influencing PMI values. A PMI value of 1 means that a Weibo narrative only occurred in one given news ecosystem. A PMI of 0 means a narrative occurred equally in each news ecosystem. If a narrative never occurred in a given ecosystem it would have an infinitely negative PMI value for that ecosystem.

This results in four, one for each news ecosystem, associated PMI values for each Weibo narrative. We take the largest of the four PMI values and consider the Weibo narrative as the most associated with that news ecosystem. Finally, we tabulate the proportion of Weibo narratives most associated with each news ecosystem.

S7 `NETINF` Algorithm

By treating each of the narrative clusters as time cascades based on when news articles and Weibo were published, we perform network inference to estimate the underlying influence network between our set of websites and Weibo. For example, given a narrative where the website Sputnik News published a news article about the narrative on January 8th, Reuters published an on January 1st, and there was a Weibo post about the narrative on January 6th, the cascade would consist of *Reuters, 1; Weibo, 6; and Sputnik News 8.*

We utilize the open-source `NETINF` algorithm to perform this network inference and determine the relative influence on Weibo by considering the weight of the inward Weibo edges in the returned network. Note that a limitation of treating narrative clusters as time cascades is that it does not account for the possibility that a narrative originated offline or in an ecosystem not measured.

Given a set of cascades C, `NETINF` aims to find the graph $\hat{G}$ that solves the optimization problem [9] where the weighted edges in the graph are connections between different websites that describe "how likely the destination website can copy or repeat information from the source website leskovec2009meme.

$$\hat{G} = \arg\max_{|G| \leq k} P(C|G)$$

where

$$P(C|G) = \prod_{c \in C} P(c|G)$$

$$P(c|G) = \sum_{T \in T(G)} P(c|T)P(T|G) \propto \sum_{T \in T(G)} \prod_{(i,j) \in T} P_c(i,j)$$

where c is a cascade and $T(G)$ is the set of all directed spanning trees on G and $P_c(i,j)$ (or the probability that the website i influenced website j in a given cascade c) is proportional to the time difference between when the two nodes publish a story about the same narrative given an exponential waiting time between article publication.

To optimize this formulation, `NETINF` considers the most likely propagation tree T for a given cascade c:

$$P(C|G) = \prod_{c \in C} \max_{T \in TG} P(c|T) = \prod_{c \in C} \max_{T \in TG} \prod_{(i,j) \in T} P_c(i,j)$$

The improvement for a graph G of the log-likelihood given a cascade c over the empty graph K is then:

$$F_c(G) = \max_{T \in T(G)} \log P(c|T) - \max_{T \in T(K)} \log P(c|T)$$

and the `NETINF` algorithm subsequently optimizes the following objective function by iteratively and greedily adding the edges with the highest marginal gain to the objective function to construct the graph $\hat{G}$:

$$F_C(G) = \sum_{c \in C} F_c(G)$$

S8 Mechanisms of Transmission

While we did not trace the exact transmission mechanisms for all Weibo posts among narratives, we observed several ways narratives make their way from overseas news websites to Weibo. First, Weibo users directly repost narratives from Russian, Ukrainian, or US news ecosystems onto Weibo, often translating the content to Chinese. For example, Figure S1 shows an account dedicated to military affairs with a large following reposting an interview clip originally published by RIA Novosti, a major Russian state-owned news agency, on Weibo. This post describes in Chinese the Russian-language news coverage of Evo Morales, the former President of Bolivia, claiming that Europe is attempting to prevent the media from reporting the truth of the war in Ukraine. In another example, Figure S2 shows a different Weibo user with a sizable following directly reposting Ukrainian-sourced news about a plane being destroyed near Kyiv.

Second, we observe foreign media outlets directly translating their content for Chinese social media. For example, Russian websites like Russia Today[S10] and Sputnik News[S11] maintain Weibo accounts and provide Chinese translations of their content on Weibo (see

Figure S1: Weibo user's direct repost of Russian-language video (Weibo username hashed).

Figure S3). The Information Office of the Ukrainian Embassy in Beijing maintains a Weibo account that regularly published Zelenskyy's remarks and pro-Ukraine statements related to the War (see Figure S3).[S12]. An example of U.S. press presence on Weibo is the official account of the Associated Press News (see Figure S3).[S13]

Third, we observe narratives in Russian, Ukrainian, or English being disseminated through Chinese-language broadcasting services and then being reposted by Weibo users. For example, as shown in Figure S4, CCTV reposted on Weibo Ukrainian news footage

[S10] https://web.archive.org/save/https://m.weibo.cn/u/6244553417?luicode=10000011&lfid=1005055991529834

[S11] https://web.archive.org/web/20241124214016/https://m.weibo.cn/u/2181597154?luicode=10000011&lfid=1005051738749010

[S12] See https://weibo.com/u/5235562548.

[S13] See https://weibo.com/u/2461865593.

Figure S2: Weibo user's direct repost of Ukrainian-sourced news sourced news about a plane being destroyed near Kyiv.

about a kindergarten being shelled in the region of Luhansk. This was then picked up by other Weibo users with one account writing "How can there be still children in kindergartens??? I really hope it is fake news. [English Translation]."'

Finally, it is important to note that narratives can traverse multiple media ecosystems before finally reaching Weibo. For example, in Figure [**?**], a news story about peace negotiations between Ukraine and Russia being at a standstill was taken from Ukrainian media before being posted by Russian media outlets, RT and RIA Novosti, and finally making its way to Weibo. In addition, other media ecosystems we do not measure may well play a role in transmission. For example, a story originally from Russian news may be picked up by Finnish news, then British news[S14], before appearing on Weibo (because British news outlets like the BBC often maintain Weibo accounts).

[S14] https://web.archive.org/web/20241125232602/https://m.weibo.cn/u/1573430940?luicode=10000011&lfid=1005051919373983&f

Figure S3: Weibo Accounts of Russia Today, Sputnik News, Ukrainian Embassy, and Associated Press.

Figure S4: Ukrainian-sourced news report about a kindergarten being shelled being picked up by China Central Television 4 and then being posted on Weibo.

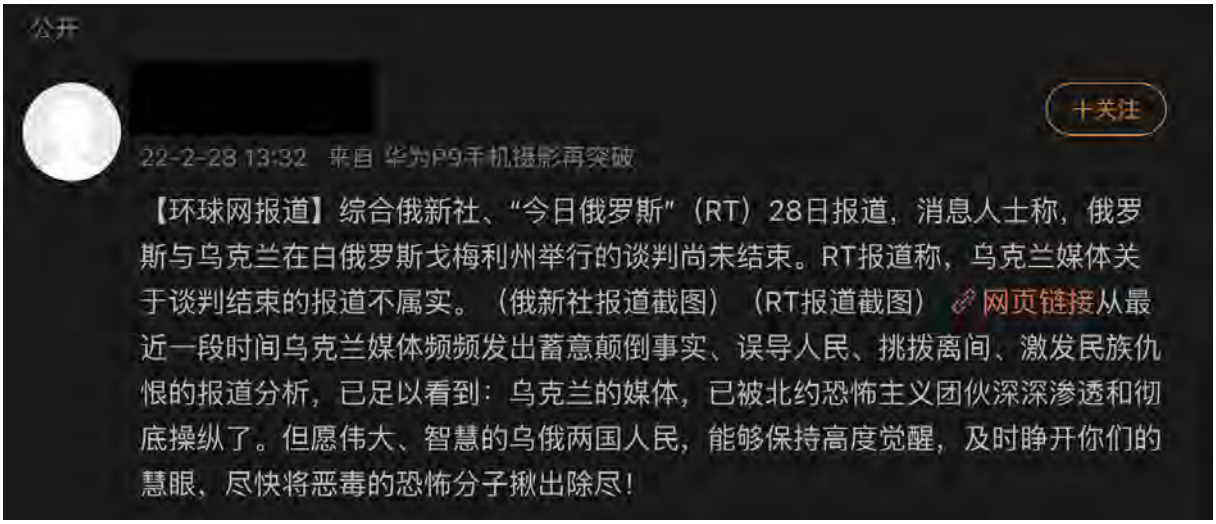

Figure S5: Story traversing from Ukrainian news outlet to Russian news outlet before being posted on Weibo.

References

[1] Alexa Internet, Inc. Top 1,000,000 Sites. http://s3.amazonaws.com/alexa-static/top-1m.csv.zip, 2022.

[2] Barret Golding. Iffy index of unreliable sources. https://iffy.news/index/, 2022.

[3] G. Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 2009.

[4] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online, nov 2020. Association for Computational Linguistics.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] O. Dinari and O. Freifeld. Revisiting dp-means: fast scalable algorithms via parallelism and delayed cluster creation. In *Uncertainty in Artificial Intelligence*, pages 579–588. PMLR, 2022.

[7] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[8] T. Gao, X. Yao, and D. Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, 2021.

[9] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):1–37, 2012.

[10] H. W. Hanley and Z. Durumeric. Partial mobilization: Tracking multilingual information flows amongst russian media outlets and telegram. *International AAAI Conference on Web and Social Media*, 2024.

[11] H. W. Hanley, D. Kumar, and Z. Durumeric. Happenstance: Utilizing semantic search to track russian state media narratives about the russo-ukrainian war on reddit. In *Proceedings of the international AAAI conference on web and social media*, volume 17, pages 327–338, 2023.

[12] H. W. A. Hanley, D. Kumar, and Z. Durumeric. A golden age: Conspiracy theories' relationship with misinformation outlets, news media, and the wider internet. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), oct 2023.

[13] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[14] X. Ming. text2vec: A tool for text to vector. https://github.com/shibing624/text2vec, 2022.

[15] S. Nagel. Common crawl's first in-house web graph. https://commoncrawl.org/blog/hostgraph-2017-feb-mar-apr-crawls, 2017. Accessed: 2022-12-29.

[16] A. Piktus, F. Petroni, V. Karpukhin, D. Okhonko, S. Broscheit, G. Izacard, P. Lewis, B. Oğuz, E. Grave, W.-t. Yih, et al. The web is your oyster–knowledge-intensive nlp against a very large web corpus. *arXiv preprint arXiv:2112.09924*, 2021.

[17] M. Post. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, page 186. Association for Computational Linguistics, 2018.

[18] A. Rondinelli, L. Bongiovanni, and V. Basile. Zero-shot topic labeling for hazard classification. *Information*, 13(10):444, 2022.

[19] K. Ruth, A. Fass, J. Azose, M. Pearson, E. Thomas, C. Sadowski, and Z. Durumeric. A world wide view of browsing the world wide web. In *Proceedings of the 22nd ACM Internet Measurement Conference*, pages 317–336, 2022.

[20] C. Samarinas, W. Hsu, and M. L. Lee. Latent retrieval for large-scale fact-checking and question answering with nli training. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 941–948. IEEE, 2020.

[21] N. Shuyo. Language detection library for java. http://code.google.com/p/language-detection/, 2010.

[22] J. Smith, H. Saint-Amand, M. Plamadă, P. Koehn, C. Callison-Burch, and A. Lopez. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, 2013.

[23] M. Szpakowski. Fake news corpus. https://github.com/several27/FakeNewsCorpus/, 2020.

[24] J. Tiedemann. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. *arXiv preprint arXiv:2010.06354*, 2020.

[25] P. D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *European Conference on machine learning*, 2001.