November 2026

ANALYSIS

Russian Disinformation on X During Platform Governance Changes

Evidence from Spring 2023

Olga Slivko[1]
Raphaela Andres[2]
Jacob Schildknecht[2]
Heiko Paulheim[3]

[1]Rotterdam School of Management, Erasmus University
[2]ZEW Mannheim
[3]University of Mannheim

Leibniz Gemeinschaft

ZEW

Russian Disinformation on X During Platform Governance Changes: Evidence from Spring 2023

Olga Slivko[1], Raphaela Andres[2], Jacob Schildknecht[2], and Heiko Paulheim[3]

[1]Rotterdam School of Management, Erasmus University

[2]ZEW Mannheim

[3]University of Mannheim

This ZEW Analysis examines the diffusion of Russian disinformation on X in spring 2023, using 1.9 million tweets and 6.5 million network interactions classified by a locally hosted Large Language Model (LLM). The study finds that disinformation spikes coincided with platform governance changes, such as the removal of restrictions on state-affiliated media accounts. It shows that disinformation is highly concentrated: a small share of source accounts originates a disproportionate share of circulating disinformation content through interconnected clusters using a limited set of narratives (e.g., delegitimization of Ukraine, elites vs. the people). Furthermore, a large share of accounts spreading disinformation were created during times of platform governance changes. The findings support the usefulness of LLM-assisted classification for large-scale monitoring and demonstrate that targeted interventions against core accounts may offer a promising approach to disrupting these dynamics. Importantly, the inclusion of user-generated content (UGC) in LLM training data raises concerns about the potential influence of false or manipulated information on AI-generated content, highlighting the importance of safeguarding both democratic discourse and the integrity of emerging AI systems.

We thank Katharina Ludwig for the data provision and many helpful discussions, and Ivan Balykin, Viktoriia Makovska, and Veronika Solopova for feedback and the expert annotations.

This project was partially funded by the German Foundation for Peace Research (DSF) in conjunction with funding provided by the German Federal Ministry of Education and Research (BMBF).

1 The Challenge of Online Disinformation

Disinformation on social media has become a major challenge for democratic societies. State actors, affiliated media outlets, and online communities can use social media to shape public opinion, undermine trust in institutions, and influence perceptions of political and geopolitical events. Russia provides an important case of such information manipulation. Since the beginning of Russias full-scale invasion of Ukraine, information campaigns linked to the Russian state have actively promoted narratives aimed at influencing perceptions of the war (European External Action Service, 2026).

Social media platforms are central channels through which these narratives spread. In response to concerns about foreign information operations, major platforms introduced measures such as labelling state-affiliated media and reducing the visibility of content linked to state-sponsored information campaigns.[1] These interventions were intended to help users better evaluate information sources and limit the reach of coordinated disinformation efforts.

In late 2022 and early 2023, X underwent substantial changes in governance, content moderation, and platform policies following its acquisition by Elon Musk. Several transparency and moderation mechanisms were modified or removed, including policies related to state-affiliated media accounts. These changes triggered public debate about the platform's ability to address misinformation, disinformation, and coordinated influence campaigns.

Understanding how disinformation evolves during periods of platform change is important for policymakers, regulators, researchers, and civil society organisations. Changes in platform governance can alter the visibility of content, the incentives of users, and the opportunities available to actors seeking to influence public debate.

This report examines the spread of Russian disinformation narratives on X during spring 2023, a period characterized by ongoing geopolitical tensions and significant platform transformations. Using a dataset of approximately 1.9 million English-language tweets related to Russia's war against Ukraine, we analyse how disinformation narratives evolved over time, which narratives were most prominent, and how they were amplified through the platform's user network.

The findings provide evidence on the dynamics of online disinformation during periods of institutional change and offer insights that can support future efforts to monitor and respond to coordinated influence operations.

[1] See for example https://www.bbc.com/news/technology-53681021

2 Data and Methodological Approach

The dataset underlying this report consists of approximately 1.9 million original English-language tweets related to Russia's war against Ukraine, collected between March 26 and May 18, 2023. Tweets were retrieved through the Twitter API using a predefined set of keywords related to the war, including references to Ukraine, the war in Ukraine, Volodymyr Zelensky, Russia, Vladimir Putin, and related topics.[2] Alongside the tweet text, metadata were gathered on user accounts, including follower counts, account creation dates, publication timestamps, and retweeting activity. For the analysis of information diffusion and amplification, the study additionally incorporates approximately 6.5 million retweets linked to the original content.

To identify disinformation content, we developed an automated classification system based on a large language model (LLM). The model was prompted to assess both the severity of disinformation and the narrative present in each tweet. Drawing on definitions from established organisations and prior research on Russian information operations, tweets were classified into four levels of disinformation severity, ranging from *no disinformation* (0) to *severe disinformation* (3). In addition, tweets were assigned to a set of twelve recurring pro-Russian narrative categories, including *delegitimization of Ukraine*, *aggressor inversion*, *critique of the West*, and related themes.

The classified data were subsequently used to examine three key dimensions of online disinformation. First, we analysed changes in the volume and composition of disinformation over time. Second, we investigated which narratives were most frequently employed and how strongly they were associated with severe disinformation. Third, we examined the network structure of content production and amplification to identify the accounts and communities most actively involved in spreading disinformation narratives.

This approach combines large-scale social media data with AI-assisted content analysis, enabling the systematic monitoring of disinformation narratives and their diffusion across the platform. A performance analysis comparing the LLM annotations with expert annotations supports our confidence in the classification. The agreement between the binary indication of *disinformation* by a majority of three experts and the LLM yielded a Cohen's κ of 0.657, conventionally interpreted as substantial (Landis and Koch, 1977, see Tables 5 and 6 for detailed analyses).

To reconstruct the network, we identified the retweets through their *RT @username:* prefix. From this, the amplifying account and the retweet text were extracted. As the Twitter API returned retweet texts only in a truncated form, we attempted to match every retweet to its classified original tweet by comparing the first 140, 125, 100, and 75 characters of the normalized texts. The most restrictive threshold was applied first. We were able to match 69% of the retweets to original tweets. Matched retweets inherit the severity level

[2]For a full list of keywords, see Table 2 in the Appendix. We note that these keywords might be biased towards pro-Ukrainian sentiment. As a result, the disinformation metrics presented in this analysis should be interpreted as conservative estimates, representing a lower bound of the actual prevalence of disinformation in the context of discussions about the war against Ukraine on X.

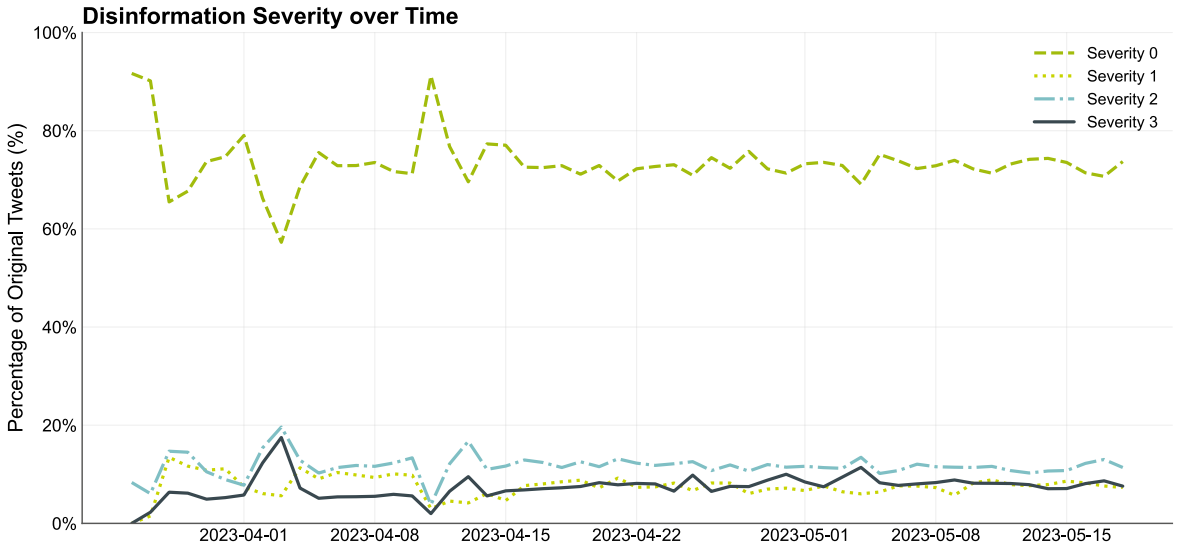
and narrative label of the original tweet and define weighted edges from amplifier to source in the network graph.

3 Results

3.1 Key Finding 1: Disinformation Spikes Coincided with Periods of Platform Governance Change

Throughout the observation period (March 26 to May 18, 2023), the share of tweets classified as disinformation increased slightly, as shown in Figure 1. Several pronounced spikes can be observed during the period. These increases were driven primarily by moderate and severe disinformation content rather than by overall increases in discussion volume.

Figure 1: Share of Disinformation Metrics over Time



Notes: This figure displays the shares of tweets containing no disinformation (green) and the shares of tweets containing low, moderate, or severe levels of disinformation over time.

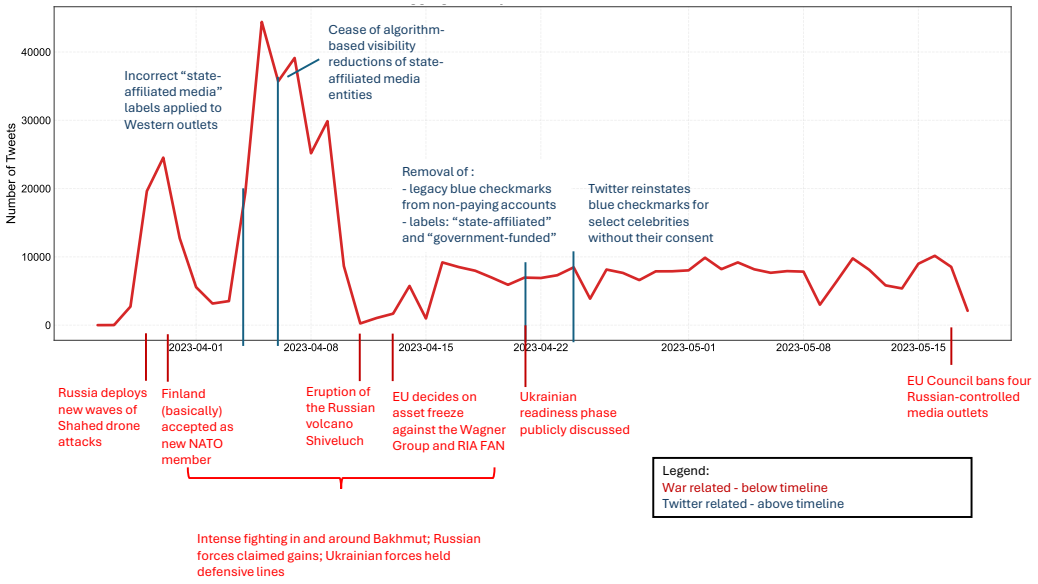
To better understand these fluctuations, Figure 2 displays the absolute number of tweets containing disinformation of severity levels 1-3 over time, alongside major platform-related, geopolitical, and other relevant events. The most prominent spike occurred in early April 2023 and coincided with significant changes in platform governance. On April 6, X removed restrictions limiting the visibility of state-affiliated media accounts, increasing their potential reach on the platform, and changed account labelling practices.[3] Second, during April 5-8, the *state-affiliated media* label was incorrectly applied to independent Western media out-

[3]https://dfrlab.org/2023/04/21/state-controlled-media-experience-sudden-twitter-gains-after-unannounced-platform-policy-change/

lets, such as National Public Radio (NPR) and the British Broadcasting Corporation (BBC).[4] This led to confusion about the interpretation of these labels among Twitter users and distorted the perceived credibility of reporting about the war.

A second notable increase occurred around March 31, which was characterized by a new wave of Shahed drone attacks by Russia[5] and Finland's accession to NATO, following the approval of Finland's membership by the Turkish parliament on March 31 and its formal accession on April 4.[6]

Figure 2: Absolute Number of Original Tweets Containing Disinformation over Time



Notes: This figure shows the development of the absolute number of tweets defined as containing disinformation of severity levels 1-3, together with events occurring during the same period. Platform specific events are shown in blue, while geopolitical and other relevant events outside the platform are shown in red.

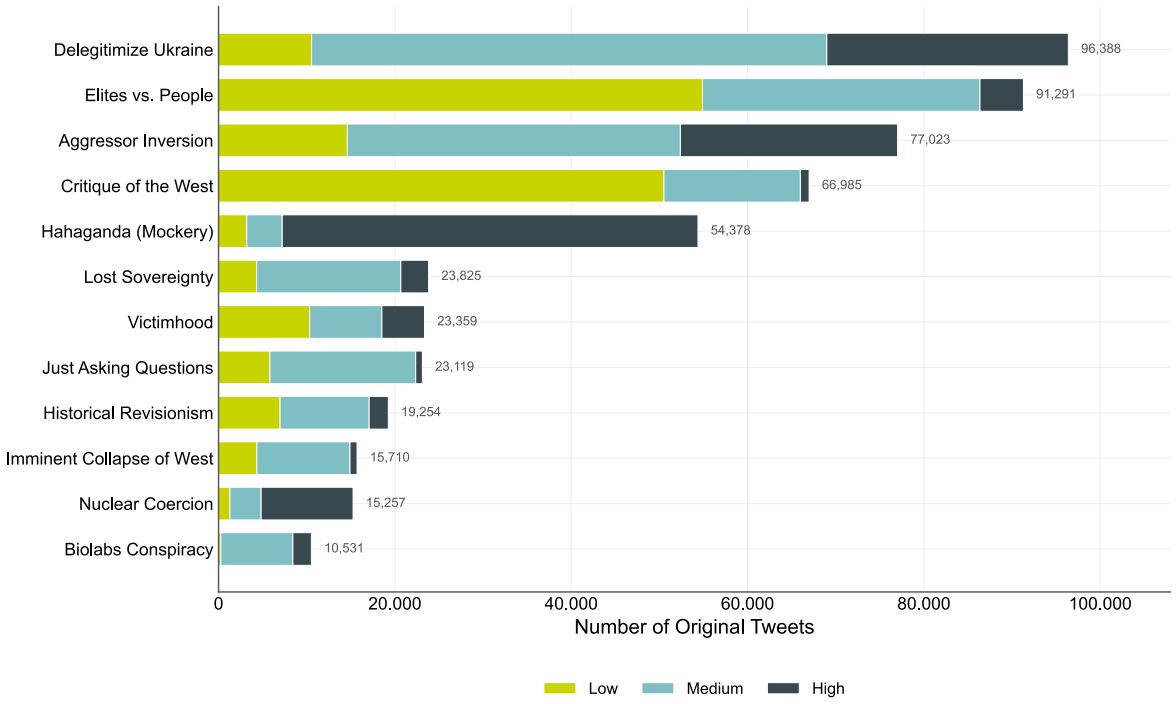
While these patterns do not establish a causal relationship, they suggest that periods of platform transition may create opportunities for disinformation actors to expand their visibility and influence. Both platform changes and external events appear to be associated with substantial fluctuations in the prevalence of disinformation.

[4] https://www.bbc.com/news/entertainment-arts-65226481
[5] https://kyivindependent.com/russia-aims-to-knock-out-ukraines-power-grid-in-new-wave-of-attacks/
[6] https://www.washingtonpost.com/world/2023/03/30/turkey-finland-nato-vote/

3.2 Key Finding 2: A Few Narratives Dominated Disinformation Content

Although the dataset contains almost two million tweets, the majority of disinformation content relied on a limited set of recurring narratives. The narratives were collected from academic research and institutional reports analysing Russian information operations and influence campaigns (EUvsDisinfo, NATO StratCom, the Euro-Atlantic Resilience Centre (E-ARC), Paul and Matthews, 2016, Amanatullah et al., 2023). Figure 3 shows the distribution of narrative categories identified in the dataset and their associated levels of disinformation severity. The most frequently observed narrative was the *Delegitimization of Ukraine*, which portrays Ukraine as a failed, corrupt, extremist, or externally controlled state. The second most common narrative, *Elites vs. the People*, depicts Western governments, media organisations, and political elites as acting against the interests of ordinary citizens. Other prominent narratives included *Aggressor Inversion*, which frames Russia as responding defensively to Western actions, and *Critique of the West*, which portrays Western institutions as hypocritical, ineffective, or morally declining.

Figure 3: Severity Composition of the Narratives



Notes: This figure displays the absolute number of tweets containing the respective narrative shown on the left of the graph. Additionally, each bar is divided by disinformation severity, ranging from low to severe.

In our dataset, the narratives differ substantially in their association with disinformation severity. While *Elites vs. the People* and *Critique of the West* frequently appear in content classified as low-severity disinformation, narratives such as *Delegitimization of Ukraine*, *Aggressor Inversion*, *Hahaganda*, and *Nuclear Coercion* are much more likely to be associated with severe disinformation.

3.3 Key Finding 3: A Small But Highly Influential Core of Accounts Initiated A Large Share of Circulating Disinformation

While many accounts participated in the amplification of disinformation content, the initiation of disinformation narratives was dominated by a small core of highly active accounts. In the subset of moderate- or severe-disinformation tweets that were retweeted at least once by accounts in our sample, we constructed the Sankey diagram in Figure 4. It reveals that a few content creators are very influential in the initiation of moderate and severe disinformation. The visualization shows the flow from the accounts posting the original tweets ("sources") on the left, through the narratives in the middle of the graph, to the accounts retweeting and amplifying these same tweets ("amplifiers") on the right. While both sides show the aggregate number of tweets and retweets, we also individually show the 15 largest sources of original tweets (i.e., accounts posting the highest number of disinformation tweets in the retweeted subset of tweets) on the left and the 15 accounts that most frequently amplify these tweets (i.e., those accounts that most often share the disinformation tweets) on the right. The numbers in parentheses indicate the number of disinformation tweets classified as at least moderate disinformation in our sample that the respective source posted or the respective amplifier retweeted.

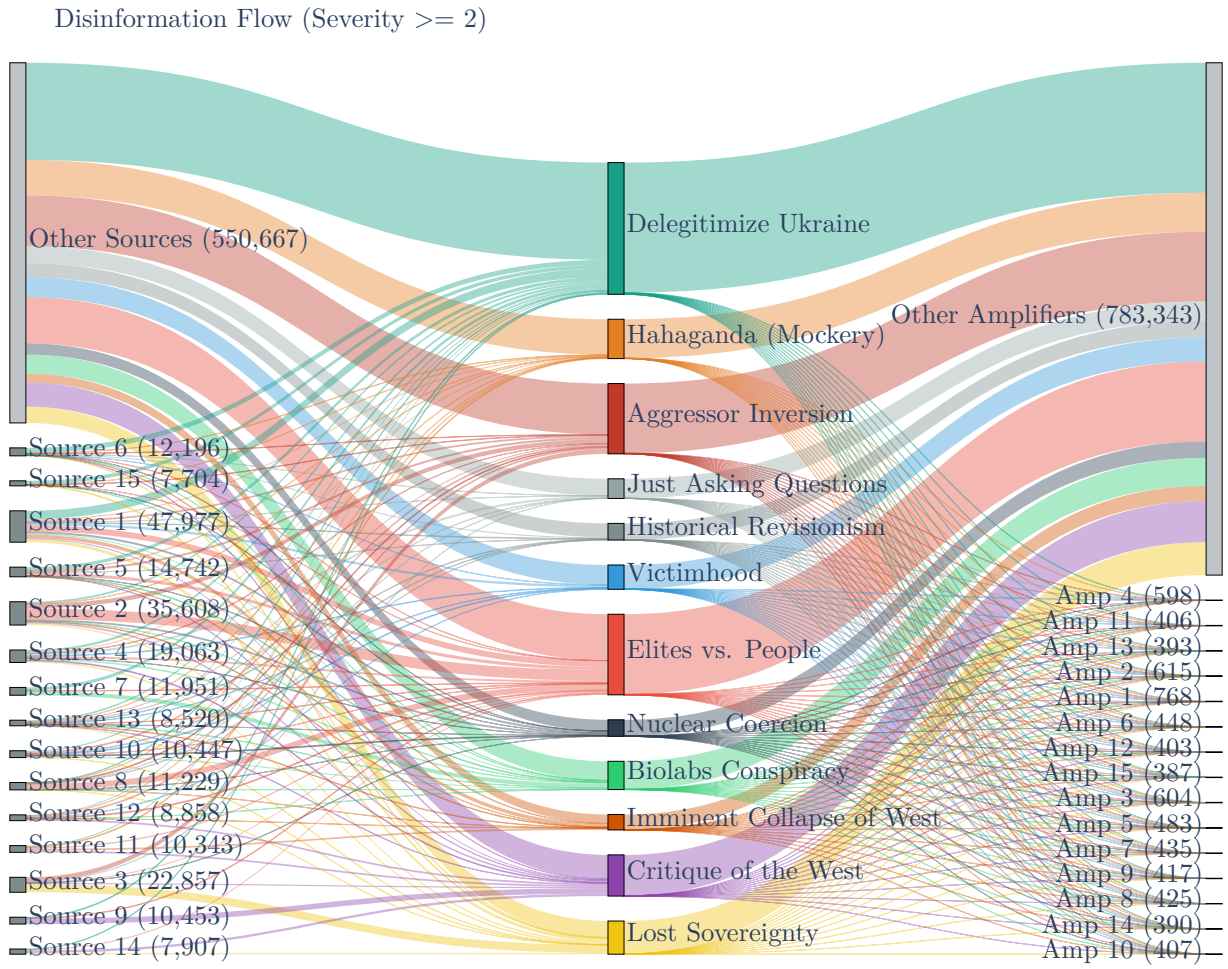
Interestingly, 30% of the volume of shared moderate and severe disinformation tweets in our sample have their origin in the top 15 source accounts, which correspond to 0.08% of all sources of shared disinformation, indicating a very high concentration in the origins of circulating disinformation. By contrast, the top 15 amplifiers, representing roughly 0.007% of all amplifiers in our matched sample, accounted for less than 1% of the moderate and severe disinformation tweets that were reshared in the sampled network. This suggests that while the origin of circulating disinformation is highly concentrated among a small number of source accounts, its amplification is considerably more dispersed across the network.

Figure 4 further shows substantial variation in the range of narratives the top sources employ: some spread multiple narratives, while others, such as sources 2 and 3, predominantly focus on a single narrative.

The analysis of the account network (see Figure 5) reveals a highly polarised and clustered network structure, which clearly separates into two distinct macro-clusters. The coloured nodes indicate the individual user accounts in our data that have at least three connections to other accounts in our data. The size of the nodes corresponds to the number of matched retweets in the sample, and the colour displays the median severity level of all tweets originating or shared by the respective account, from green (severity level 0) to red (severity level 3). The first cluster is characterized by a high concentration of Severity 0 (no disinformation) accounts, shown in the upper left corner. The second cluster, shown in the lower right corner, serves as the primary hub for accounts spreading disinformation narratives.

These clusters exhibit high intra-connectivity but are only partially linked to one another, resulting in a

Figure 4: Sankey Diagram of Main Spreaders of Severity 2 and 3 Disinformation

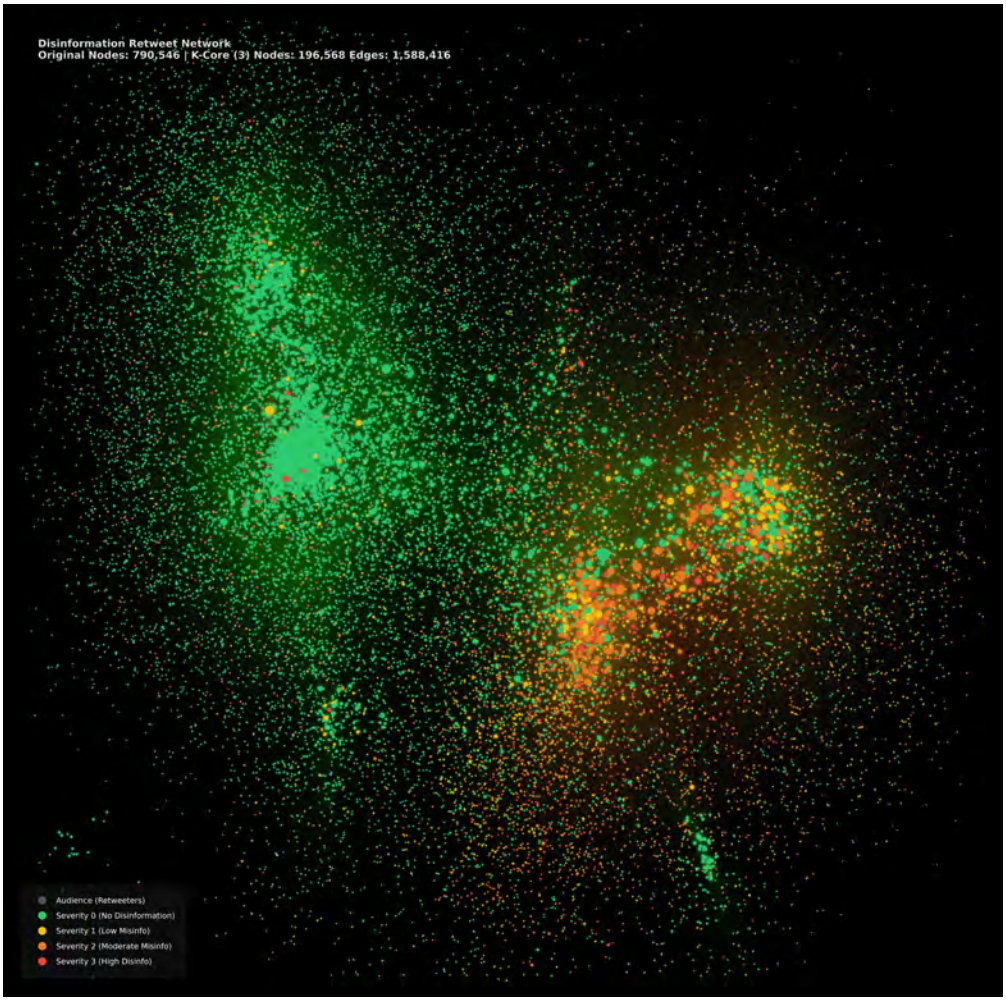


Notes: This figure displays the producers of shared disinformation of Severity 2 and 3 on the left and the main amplifiers on the right. The account names are anonymised due to data protection reasons. The middle part of the figure indicates the narrative that the respective tweet was assigned to.

modularity score of 0.6306 across 12,770 communities. Notably, the PageRank analysis shows that the network's core is dominated by high-severity disinformation, as 7 of the top 10 most central nodes are classified as Severity 3 accounts.

Taken together, the network analysis reveals that disinformation is concentrated within a separate cluster. However, this cluster is highly influential, as its high-severity disinformation accounts also occupy central positions in the overall network.

Figure 5: Network Graph



Notes: This figure displays the network of accounts in the sample. The size of the nodes corresponds to the number of retweets in the sample, while node colour indicates the median disinformation severity of an account's tweets, ranging from 0 (no disinformation; green) to 3 (severe disinformation; red).

3.4 Key Finding 4: Accounts Associated with Disinformation Were Younger and Emerged During Platform Transitions

Accounts associated with severe disinformation differ systematically from the rest of the information ecosystem. Table 1 shows that accounts predominantly sharing severe disinformation (Severity 3) were younger,[7] had fewer followers, and were more likely to have been created after the start of Russia's full-scale invasion of Ukraine in February 2022.

On average, accounts associated with severe disinformation were approximately one year younger than

[7]The age of an account is measured by its creation time. For accounts associated with a natural person, it must not be confused with that person's age.

accounts sharing no disinformation. They also possessed substantially smaller follower bases, with a median follower count of 194 compared to 300 for accounts associated primarily with non-disinformation content. At the same time, accounts spreading severe disinformation were more likely to have been created after the outbreak of the war, suggesting a greater reliance on newer accounts.

Table 1: Account Metadata by Severity Level

Severity	Count	Avg Age (years)	Median Followers	Post-War %
0	544,131	6.66	300	18.3%
1	87,730	6.03	259	24.6%
2	103,844	5.78	244	24.9%
3	54,102	5.57	194	26.5%

Notes: This table compares account statistics by severity level. The severity level is defined based on the median severity level of all tweets by each account.

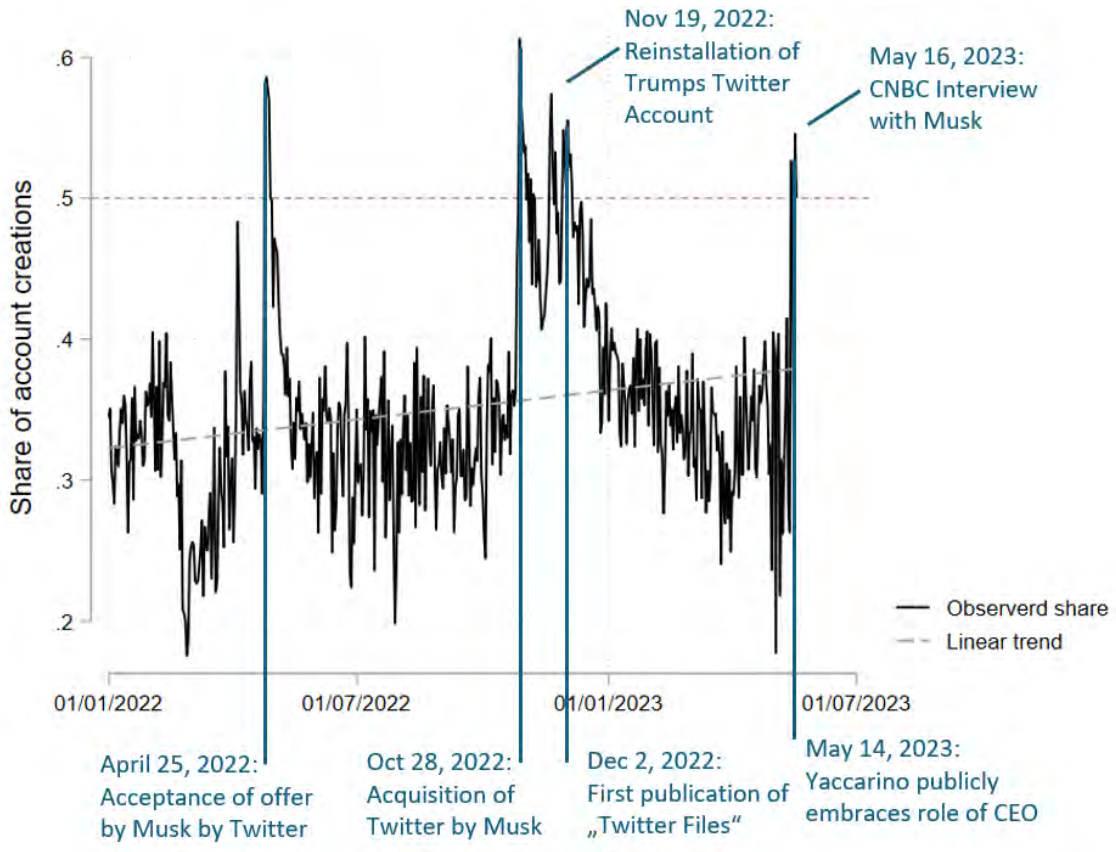
To investigate this further, we plotted the share of all newly created accounts that were subsequently classified as disinformation-spreading accounts (i.e., accounts with a median severity above 0) over time (see Figure 6). This plot reveals several spikes during which more than half of the newly created accounts are identified as spreaders of disinformation. All of these dates coincide with major platform events, as shown in the figure. The largest spike of newly created disinforming accounts occurred on October 28, 2022, the day Twitter was officially acquired by Elon Musk. On that date, more than 60% of newly created accounts were subsequently classified as disinformation-spreading accounts (i.e., accounts with a median severity above 0), corresponding to 2,415 accounts. This represents an almost 60-fold increase over the average daily number of such account creations observed (41.9 accounts per day in our dataset). The temporal correspondence between these registration spikes and platform milestones may be consistent with broader strategic expansion of disinformation campaigns.

While these patterns cannot establish intent or coordination, they indicate that periods of platform transition coincided with unusually high levels of account creation among users who subsequently became active spreaders of disinformation. These findings suggest that platform governance changes may be associated not only with the visibility of disinformation but also with the composition of the accounts participating in its dissemination.

4 Policy Implications and Recommendations

The evidence underscores that social media platforms provide critical infrastructure in the information ecosystem. When content moderation is weakened and uncertainty about platform governance increases, the spread of harmful content may increase.

Figure 6: Creation of Disinformation Spreading Accounts



Notes: This figure displays the shares of account creation dates of accounts with a median disinformation severity above 0.

The analysis of the disinformation flow in this study suggests that a limited number of source accounts generated a large share of moderate and severe disinformation tweets that contaminated a substantial cluster in the network. This means that targeted content moderation, for example, removing these high-volume source accounts that are especially active in the production of disinformation, may offer an efficient approach to disrupting these dynamics.

These findings are relevant to platform managers, policymakers, and the general public, especially in light of the growing reliance on LLMs. The inclusion of user-generated content (UGC) in LLM training data raises concerns about the potential influence of false or manipulated information on AI-generated content. This may cause AI systems to reproduce and amplify disinformation present in their training data.[8] To safeguard both public discourse and emerging technologies, platform policies must prioritise consistent, transparent moderation, and regulatory frameworks should support targeted enforcement and data integrity standards.

Finally, our findings also highlight the importance of researchers' access to platform data. This study was conducted before X discontinued free data access for researchers and replaced it with a paid model.

[8] https://www.newsguardtech.com/special-reports/anthropic-ai-chatbot-claude-russia-iran-propaganda/

Conducting this study with today's pricing model would have cost more than USD 50,000. To effectively monitor and analyse social media, and given the important role of social media for public discourse, it is paramount that researchers can obtain access to social media data on reasonable terms.

Appendix

Data Extraction

Tweet collection was conducted using the Tweepy Python package (Roesslein, 2009), which provides a structured interface for accessing the Twitter API.

Table 2: Distribution of Tweets by Keyword

Keywords Ukraine War	**Number of distinct tweets retrieved**	**Share of retrieved tweets (%)**
stoprussianagression	551	0.008
putinhitler	1,928	0.028
ukrainecrisis	2,010	0.029
ukraineinvasion	3,802	0.056
ukraineunderattack	4,864	0.070
fckputin	14,838	0.217
russianukrainwar	28,033	0.425
stoprussia	57,505	0.842
ukrainerussianwar	63,706	0.931
ukraincrisiswar	117,887	1.725
ukrainewar	149,767	2.110
standwithukraine	226,909	3.319
Kiev	339,906	4.045
Zelensky	570,260	8.337
Putin	1,042,603	15.721
Russian	1,144,267	16.775
Russia	2,044,839	29.960
Ukraine	2,059,911	30.200

Notes: Since tweets can contain multiple keywords, the counts and percentages are not mutually exclusive.

Description of Russian Narratives

The narratives defined in this study are based on the following sources: EUvsDisinfo, NATO StratCom, The Euro-Atlantic Resilience Centre (E-ARC), Paul and Matthews, 2016, Amanatullah et al., 2023.

Table 3: Pro-Russian Narratives

Letter	Narrative	Definition
A	Elites vs. the People	Claims that Western elites or media are corrupt, hypocritical, hiding the truth, or oppressing ordinary people. This narrative portrays Western leadership as self-serving and disconnected from citizens, often in contrast to Russia's interests. Assign A only if the tweet specifically criticises Western elites, governments, or media as deliberately sacrificing, deceiving, or exploiting ordinary people. Do not assign A when the criticism is directed at Russia, Russian leadership, the Kremlin, or pro-Russian figures, or when it is merely insulting an individual without reinforcing a pro-Russian framing.
B	Critique of the West	Claims that the West is morally decaying (e.g., LGBT, multiculturalism), anti-family, and hypocritical. Emphasises Western double standards, fake democracy, and portrays global institutions (UN, EU, NATO, World Bank) as tools of domination.
C	Victimhood	Claims that Russia is a moral actor and a victim of Western aggression or unfair criticism. Russia is framed as defending peace, Orthodox values, or global stability. Assign C only if the tweet depicts Russia or Russians as innocent, unfairly targeted, misunderstood, or morally superior. Do not assign C when the tweet condemns Russia as an aggressor, criminal, or barbaric, or when it calls for Russia to be punished or isolated.
D	Lost Sovereignty of countries other than Russia	Claims that countries like Ukraine or EU members are not truly sovereign, but Western puppets.

Continued on next page

Continued from previous page

Letter	Narrative	Definition
E	Imminent Collapse of EU, NATO, or West	Claims that EU, NATO, or West is about to fall apart economically, culturally, or politically. Suggests Western sanctions harm Europe more than Russia. Blames energy or food crises on EU/NATO.
F	Hahaganda (Mockery & Ridicule) that supports pro-Russian narratives	Sarcastic, mocking tone targeting Ukraine, Western leaders and institutions, or war victims. Assign F only if sarcasm/mockery targets Ukraine, Western leaders/institutions, or war victims in a way that reinforces pro-Russian templates. If the mockery is directed at Russia, Russian leadership, army, propaganda, or pro-Russian activists, then classify as: "label": 0, "narrative": "NA".
G	Delegitimizing Ukraine	Ukraine is portrayed as a failed state, fascist regime, Nazi, terrorist, or (Western) puppet state. Western military or humanitarian aid is framed as illegal, wasted, or supporting corruption/extremism. Assign G only if the tweet explicitly portrays Ukraine in these negative terms. Do not assign G when the tweet rejects, ridicules, or argues against these claims, even if the words corrupt or failed appear as part of a rebuttal.
H	Weapons & Biolabs Conspiracies in Ukraine	Promotes discredited claims of Western-origin biological or chemical weapons in Ukraine, organ harvesting, Nazi ties of Ukraine, or other hidden agendas.
I	Just Asking Questions Strategy	Disinformation framed as neutral inquiry or open debate (e.g., Could it be that NATO knew all along?). Assign I only if the tweet explicitly uses rhetorical questions or speculation to cast doubt on established facts or to shift blame toward the West, Ukraine, or other actors in a way that overlaps with pro-Russian narratives. Do not assign I when the tweet makes direct assertions, reports concerns as facts, or expresses personal opinion without using a just asking style.

Continued on next page

Continued from previous page

Letter	Narrative	Definition
J	Aggressor Inversion	Claims that Russia was forced to act defensively against NATO expansion or Western encirclement, portraying Russian aggression as justified or pre-emptive self-defence. It often suggests that NATO/West provoked Russia, Russia is defending itself, Ukraine attacked first, or blames Ukraine/West for civilian casualties. Assign J only if the tweet justifies or excuses Russian aggression by shifting blame to Ukraine or the West. Do not assign J when the tweet explicitly condemns Russia as the aggressor, criticises Russian actions, or describes Russian attacks as unprovoked.
K	Historical Revisionism / Great Power Status of Russia	Claims that Western countries are trying to diminish Russia's historical achievements or territorial claims. Frames current conflicts as a continuation of historical grievances.
L	Nuclear Coercion / Red-Line Setting	Uses threats of extreme escalation (e.g., nuclear weapons, WWIII) to coerce the West. This narrative frames Western support for Ukraine as a dangerous provocation that crosses a "red line" and forces Russia to consider a catastrophic response. The goal is to shape Western decision-making and public opinion by inducing fear, thereby limiting or delaying military aid to Ukraine.

Table 4: Sample tweets per narrative and severity level

Narrative	Sev.	Example Tweets
Elites vs. People	1	You have no business speaking on any kind of ETHICS you and your phony Russia dossier, espionage(spy) against a setting president , Was that part of Obama shadow government & or his campaign? And you wanna talk about ethics , Never forget Benghazi !
Elites vs. People	2	Its Politico. Its American media. Any little bit of truth is there to make the lies go down more smoothly. And the US has nukes all around Russia.

Continued on next page

Continued from previous page

Narrative	Sev.	Example Tweets
Elites vs. People	3	Finland was admitted to NATO in record time. Ukraine stays outside. The Ukies who still defend NATO with their bodies don't seem to realize they're sacrificial lambs of the military alliance and go on marching to be slaughtered. If this is a people, what's definition of a herd?
Critique of the West	1	He made the same ridiculous statements about Crimea. Zelensky believes that in Europe, as well as in Ukraine, there is a total information vacuum.
Critique of the West	2	With the power of hindsight. We know that #China has profited the most from #NordstreamSabotage . They have managed to ensure #Russia will be seen as a villain by the west (they were already, but now they are iridimable), thus making Russia more dependent on #Yuan. #WeKnow.
Critique of the West	3	I LOVE Putins speeches & how he quotes scripture. I love how he talks about about Gd, Christian values,& Jesus. I love his alliance.I love how he talks about Conservatism &national pride. I love how he calls evil we face SATANIC &DEMONIC. I love how he calls out De3p Stat3
Victimhood	1	When Russia wins, every Church needs to be rebuilt, and better than ever.
Victimhood	2	Putin will help to save the West.
Victimhood	3	Wow thats so cringe. Russia defeated Nazism and pushed Hitler back to Berlin. Russias sacrifice is what Europe should be thankful for. Zelensky is trying to erase Slavic culture in Ukraine and Europeanize. Why? Pure greed. Thats why he meets with billionaire bankers. He wants
Lost Sovereignty	1	The guy you voted for just agreed to send them another $2.6B, they arent getting rid of them theyre funding them. I dont entirely agree with Russia but I dont give a flying fuck if they take Ukraine.

Continued on next page

Continued from previous page

Narrative	Sev.	Example Tweets
Lost Sovereignty	2	As longs as a US tax payer Im financing this nonsense then Im also entitled to opine on the resolution. I would not think twice but split Ukraine in half and let Russia keep the East. Brazil 1) is not racketeering from US tax payers 2) doesnt have borders w Russia 3) has no
Lost Sovereignty	3	Except they feed people into a meat grinder for optics, clinging to thread a bakmut, not letting them retreat and allowing russia to just shell them to death. Zelensky has no heart for the Ukrainians, he is US puppet, the proxy war is a US opportunity
Imminent Collapse of West	1	The conquest of Russia by the West has driven Russian money to China, giving a serious boost to the Chinese economy. If the trend persists, and middle-east joins in, this is the final nail in the coffin of the American superpower.
Imminent Collapse of West	2	I was talking about USA. But now that you mention it....Ukraine will never win a war against Russia. Every western country has sent $$ and weapons, we've sanctioned Russia so much WE have destroyed west economies. Ukraine was supposed to win in 3 months haha. A year later..
Imminent Collapse of West	3	NATO to arm Ukraine in a war against Russia. A foolish notion which will lead to the end of Ukraine and the end of NATO
Hahaganda (Mockery)	1	So they can't intercept a slow flying drone (supposedly) all the way from ukraine, so you think they'll take their chances with NATO? That's funny
Hahaganda (Mockery)	2	Look his source is Kiev independent, lol
Hahaganda (Mockery)	3	Between Putin & Obi who is a killer? If you cannot answer honestly, I wonder what else you could.
Delegitimize Ukraine	1	Radio 3 Breakfast had music from the countries of the 26 finalists. Proper music but nevertheless planting the Eurovision seed! No Zelensky because they want to avoid politics! That's exactly why Ukraine won last year!

Continued on next page

Continued from previous page

Narrative	Sev.	Example Tweets
Delegitimize Ukraine	2	Your NOT doing a dam thing but spending billions of our tax payers dollars in Ukraine to pay for guns and ammo to kill people everyday good job!!!! You of ALL people should know there are consequences for not following the rules like these representatives did in Tennessee.
Delegitimize Ukraine	3	EU is run by children and grandchildren of nazis and colaborators with nazis pretending to deny their origins. Their goal is seeking revenge against Russia, but the grandchildren of Red Army are decimating adorers of genocidal nazi Stepan Bandera, current USA-NATO proxies.
Biolabs Conspiracy	1	Depleted Uranium is not radiactive. Its different kind of metal..Russia been using DU for its T90s..it wont even explode .
Biolabs Conspiracy	2	NATO UNDERGROUND KIEV MEETINGS HAS BEEN DISTROY BY ussRussia no one talk about it only journalists and closed circuit news knows ..Zelenzky was supposed to be at that meeting....
Biolabs Conspiracy	3	Unconfirmed reports that Wagner have cut adrenochrome supply lines with destruction of Bakhmut bio labs. Clearly, Zelensky's body and mind are quickly disintegrating. How does this make you feel? Be True to Yourself!
Just Asking Questions	1	Someone help me out, how are they paying for the imports? I though Russia was out of Swift. Or are they exchanging goods?
Just Asking Questions	2	You are going after Russia, but what I want to know is if Zelensky informants gave up the location to munitions. Got a gut feeling. You should probably look into that instead.
Just Asking Questions	3	This and that Zelensky left Ukraine the same day of the attack are clear evidence that he knew about it. Also he flew with a US plane. So the US knew as well. Its same as with Kerch bridge attack and Nordstream. Anyone with an intact brain knows who was behind.

Continued on next page

Continued from previous page

Narrative	Sev.	Example Tweets
Aggressor Inversion	1	I'm not a huge fan of Russia, I'm a veteran of the Cold War. That said, what do you think the US would do if Mexico signed a defense pact with China and then allowed China to set up nuclear bases within range of the US border? Nations don't go to war because it's fun.
Aggressor Inversion	2	Had it been that RUSSIA is not well prepared for what is to come they would have been leveled to the ground by WEST. The question for AFRICA is . Are we thinking?
Aggressor Inversion	3	Defend freedom in Europe. Dismembering NATO can be positive first step towards future peace & freedom. Disarming & dismantling the Zelensky regime in Kyiv, should be immediate action, to stop Ukrainians dying.
Historical Revisionism	1	One of history's great ironies is that during the "Cold" War, the Russians somehow browbeat the pathetic UN to allow there to be separate representation for Belarus (then known as "Byelorussia") &... Ukraine. Yup – as Yogi said, "You could look it up" The Soviets had 3 votes.
Historical Revisionism	2	He thinks that it's not his war, his goal - to prevent nuclear disaster. But he's so wrong... Why? Because, as putin sad, the borders of russia does end nowhere. And, i bet, he won't like to live in rf
Historical Revisionism	3	Yes, but the plan is to break up Russia into insignificant parts that can be managed so that those resources are in the hands of the owners of the planet. Although it won't work out that way Finland is thinking it could.
Nuclear Coercion	1	What is going on? This is somewhat concerning. Is this military leadership in the air due to a nuclear concern? Hope russia isnt thinking of pulling something here.
Nuclear Coercion	2	Moldova next if he wins this. Then ex Eastern bloc countries. Pray to God Ukraine wins. If they don't you'll really see NATO and Russia square off. That will involve nukes. Western leaders are not stupid. Ukraine is the only thing stopping this spilling over into wider conflict.

Continued on next page

Continued from previous page

Narrative	Sev.	Example Tweets
Nuclear Coercion	3	Putin, nuke Chitown first.

Performance of the LLM against Expert Annotations

To assess the reliability of the LLM and the classified data, we asked three experts in the field[9] to manually classify a random subset of 100 tweets into the four disinformation severity levels from 0 to 3.

Comparing the severity scores of the three experts, we acknowledge that assigning severity levels is a difficult task, as the inter-rater agreement yielded a Cohen's κ of 0.43, which is considered "moderate" (Landis & Koch, 1977). Therefore, we focus on binary indicators for the presence of disinformation (severities 1, 2, and 3) according to each of the experts, the LLM, and the majority of experts. The *expert majority indicator* is derived from the three expert annotations and equals 1 if at least two of the three experts defined a tweet as containing any severity of disinformation.

Table 5 shows the pairwise Cohen's κ of the binary indicator for the presence of disinformation between the individual experts, the expert majority, and the LLM. The automated classification performed by the LLM shows "substantial agreement" with the expert annotations, based on Cohen's κ of 0.657 with the majority expert rating (Landis & Koch, 1977).

Table 5: Pairwise Cohen's κ

	Model	Expert_Majority	Expert1	Expert2	Expert3
Model	1	.6570122	.6570122	.5807927	.4730679
Expert_Majority	.6570122	1	.8582566	.8582566	.6899662
Expert1	.6570122	.8582566	1	.7165131	.5772266
Expert2	.5807927	.8582566	.7165131	1	.5772266
Expert3	.4730679	.6899662	.5772266	.5772266	1

Notes: The table shows all pairwise Cohen's κ of the binary classifications into *No Disinformation* and *Disinformation* between all three expert annotators, the composed expert majority and the LLM.

Table 6 provides further evidence that the LLM performs well compared to the majority of experts, with particularly high scores for tweets classified as *No Disinformation*: most cases classified as non-disinformation are aligned with the expert majority and almost all non-disinformation cases according to the expert majority were captured.

An analysis of the tweets classified as *Disinformation* by the majority of experts shows that the LLM appears relatively conservative in detecting Russian disinformation. The recall of 0.65 indicates that of all

[9]Prof. Dr. Ivan Balykin, Viktoriia Makovska, Dr. Veronika Solopova

Table 6: Classification Report (binary: 0 vs. 1+)

	Precision	Recall	F1-score	Support
0-None	.93	.97	.95	83
1-Disinformation	.79	.65	.71	17
Accuracy	.	.	.91	100
MacroAvg	.86	.81	.83	100
WeightedAvg	.91	.91	.91	100

Notes: The table summarizes performance metrics of the binary classifications of the LLM into *No Disinformation* and *Disinformation* compared to the majority rating of all three expert annotators.

the tweets that have been classified as *Disinformation* by the majority of experts, the model defines only around two-thirds as *Disinformation*.

System Prompt

Procedure

To classify the data by disinformation severity and narrative used in the tweets, we designed a structured prompt.

We started by collecting trustworthy definitions of disinformation and operationalised it using a four-level scale ranging from 0 (no disinformation) to 3 (strong disinformation). We added the reasoning as a model output to better understand the potential pitfalls of the definitions and to identify and resolve ambiguities. We iteratively refined the prompt, with three members of the research team reviewing the resulting labels and reasoning. We also added example tweets to guide the model.

After achieving a good classification performance, we proceeded to narrative classification. Narrative definitions are based on Amanatullah et al., 2023, EUvsDisinfo, NATO StratCom, The Euro-Atlantic Resilience Centre, and Paul and Matthews, 2016. We again conducted several iterations to refine the boundaries between the narratives and specified that no narrative should be assigned when the disinformation severity is 0.

We explicitly specified the internal logic to ensure that all steps of the classification are executed.

All tasks were conducted using Qwen 3 80B (temperature = 0.0, top_p = 1.0).

Prompt

You are a classification model that evaluates whether a tweet promotes or reflects one or more of the pro-Russian disinformation narratives using the definitions and pro-Russian narrative templates below. Respond

with a single, raw JSON object.

The JSON object must contain three keys:

1. "label": A number from 0 to 3 indicating the strength of alignment with any of the Pro-Russian narrative templates according to the Disinformation Scale.
2. "narrative": A letter from A to L showing which of the pro-Russian narrative templates is dominant in the tweet. If there is no narrative present, set the value to "NA"
3. "reasoning": A brief, one-sentence explanation for your classification.

Do not add any text or explanations before or after the JSON object.

Definitions (EUvsDisinfo, NATO StratCom, The Euro-Atlantic Resilience Centre (E-ARC), Paul & Matthews, 2016).

- Disinformation: Verifiably false or misleading information, shared deliberately to influence opinion or obscure truth.
- Propaganda: "An intentional effort to shape the relationship between an individual target of information (the subject) and their environment (the object) by relying on the dissemination of symbolic meaning in order to support a particular course of the subject's activity in relation to specific objects of activity." (Asmolov, 2021)
- Russian disinformation often combines ideological narratives with manipulative rhetorical tactics to support Russian geopolitical aims.

Pro-Russian narrative templates:

A. Elites vs. the People: Claims that Western elites/media are corrupt, hiding the truth, or oppressing "ordinary people."
B. Critique of the West: Claims that the West is morally decaying (e.g., LGBT, multiculturalism), anti-family, and hypocritical. Emphasises Western double standards, fake democracy, and portrays global institutions (UN, EU, NATO, World Bank) as tools of domination.
C. Victimhood: Russia portrays itself as a moral actor, victim of Western aggression or unfair criticism. Russia is framed as a protector of peace, Orthodox values, or global stability.
D. Lost Sovereignty of other countries: Claims that countries like Ukraine or EU members are not truly sovereign, but Western puppets.
E. Imminent Collapse: Claims that EU, NATO, or West is about to fall apart economically, culturally, or politically. Suggests Western sanctions harm Europe more than Russia. Blames energy or food crises on EU/NATO.

F. Hahaganda (Mockery & Ridicule): Sarcastic, mocking tone targeting Ukraine, Western leaders and institutions, or war victims.

G. Delegitimizing Ukraine: Ukraine is portrayed as a failed state, fascist regime, a Nazi, terrorist, or (Western) puppet state. Western military and humanitarian aid framed as illegal, supporting criminal or extremist actors, or is wasted due to corruption in Ukraine.

H. Weapons & Biolabs Conspiracies: Promotes discredited claims of Western-origin biological or chemical weapons in Ukraine, organ harvesting, Nazi ties, or hidden agendas.

I. "Just Asking Questions" Strategy: Disinformation framed as neutral inquiry or "open debate" (e.g. "Could it be that NATO knew all along?").

J. Aggressor Inversion: Claims that Russia was forced to act defensively against NATO expansion or Western encirclement. It often portrays Russian aggression as justified or pre-emptive self-defence. "NATO/West provoked Russia", "Russia is defending itself", "Ukraine attacked first." Accuses Ukraine of genocide or staging atrocities. Claims Russia is falsely blamed. Civilian casualties in Ukraine are portrayed as caused by Ukraine or the West.

K. Historical Revisionism / Great Power Status: Claims that Western countries are trying to diminish Russia's historical achievements or territorial claims. Frames current conflicts as a continuation of historical grievances.

L. Nuclear Coercion / Red-Line Setting: Uses threats of extreme escalation (e.g., nuclear weapons, WWIII) to coerce the West. This narrative frames Western support for Ukraine as a dangerous provocation that crosses a "red line" and forces Russia to consider a catastrophic response. The goal is to shape Western decision-making and public opinion by inducing fear, thereby limiting or delaying military aid to Ukraine.

Examples: { EXAMPLE_TWEETS }

Steps (Internal Logic You Should Follow):

1. Read the tweet carefully.
2. Identify if the narrative in the tweet aligns with any pro-Russian narrative template.
3. If it aligns with a pro-Russian narrative template, select the narrative that dominates and determine the strength of alignment using the Disinformation Scale. If no alignment with narrative detected, return 0 for label according to Disinformation Scale and NA for the narrative according to Pro-Russian narrative templates.

Disinformation Scale:

0 - None: No disinformation or pro-Russian narratives present. Tweet is factual or neutral. Pro-Ukrainian sentiment or opposition against Russia is not disinformation and should be classified as 0.

1 - Low: Slight rhetorical overlap with a pro-Russian narrative.

2 - Moderate: Clear alignment with a pro-Russian narrative or misleading framing; may indirectly support pro-Russian goals.
3 - High: Strong disinformation or propaganda; explicitly promotes a pro-Russian narrative.

Task:
Given the tweet below, return a single number: 0, 1, 2, or 3 according to the Disinformation Scale for "label", return a letter of the dominating pro-Russian narrative from the list of Pro-Russian narrative templates (or "NA" for "label" 0) for "narrative", and provide a brief reasoning for the classification. Output ONLY the raw JSON object, nothing else.

Tweet: {text_to_classify}
Answer:

References

Amanatullah, S., Balani, S., Fraioli, A., McVicker, S. M., & Gordon, M. (2023). Tell us how you really feel: Analyzing pro-kremlin propaganda devices & narratives to identify sentiment implications. *The Propwatch Project. Illiberalism Studies Program Working Paper. https://www. illiberalism. org/tell-us-how-you-really-feel-analyzing-pro-kremlinpropaganda-devices-narratives-to-identify-sentiment-implications*.

European External Action Service. (2026). *4th EEAS Report on Foreign Information Manipulation and Interference Threats* (tech. rep.) (https://www.eeas.europa.eu/eeas/4th-eeas-annual-report-foreign-information-manipulation-and-interference-threats_en). European Union.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Paul, C., & Matthews, M. (2016). The russian firehose of falsehood propaganda model. *Rand Corporation*, *2*(7), 1–10.

Roesslein, J. (2009). Tweepy documentation [Online] http://tweepy. readthedocs. io/en/v3].

ZEW

Impressum

Herausgeber: ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung
L 7, 1 · 68161 Mannheim · Deutschland · info@zew.de · www.zew.de ·
Tel.: +49 (0) 621/1235 01
Präsident: Prof. Dr. h. c. Achim Wambach, PhD · Kaufmännische Geschäftsführerin: Claudia von Schuttenbach

Anmerkung zum Zitieren aus dem Text: Es ist gestattet, Auszüge aus dem Text in der Originalsprache zu zitieren, insofern diese durch eine Quellenangabe kenntlich gemacht werden.
© ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim