Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Fear and Loathing on the Frontline: Decoding the Language of Othering by Russia-Ukraine War Bloggers

Patrick Gerard[1], **Tim Weninger**[2], **Kristina Lerman**[1]

[1]Information Sciences Institute, University of Southern California, Marina del Rey, California, United States
[2]University of Notre Dame, Notre Dame, Indiana, United States
pgerard@isi.edu, wtheisen@nd.edu, tweninger@nd.edu, lerman@isi.edu

Abstract

Othering—the process of portraying an outgroup as fundamentally different and inferior—often escalates into framing the outgroup as an existential threat, thereby legitimizing exclusion and violence. Throughout history, othering has played a central role in conflicts, from genocides in Nazi Germany and Rwanda to contemporary hostility toward migrants in the US and Europe. Traditional computational methods, such as those used for hate speech detection, frequently overlook the subtle, context-dependent nature of othering language, limiting their effectiveness in real-time detection and analysis. Our work addresses these limitations through three key contributions: (1) a computational framework that combines sociological theory with large language models (LLMs) to identify and analyze othering language, (2) an in-depth examination of othering discourse dynamics, focusing on attention patterns and its interplay with moral framing, and (3) a rapid domain adaptation enabling robust analysis across different platforms and contexts. We apply our framework to a large corpus of Telegram messages from Russo-Ukrainian war bloggers and political discourse on Gab, revealing several previously unquantified patterns: othering rhetoric surges during crises, often intertwines with moralized language, and escalates during critical periods. Our findings demonstrate that this approach not only surpasses existing hate and fear speech detection methods but also offers actionable insights for anticipating and mitigating threats to social cohesion in conflict-prone environments.

Code — https://github.com/patrikgerard/othering_language

Introduction

In crises, people rely on social media for information, making these platforms powerful shapers of public perception (Tsoy et al. 2021). This creates opportunities for manipulation through propaganda, particularly through rhetoric that portrays certain groups as dangerous and separate, which can fuel radicalization and conflict (Saha et al. 2021; Cervone, Augoustinos, and Maass 2021). Such rhetoric ranges from explicit hate speech and dehumanization (Kennedy et al. 2023; Buyse 2014) to subtler forms of fear speech (Saha et al. 2023; Schulze, Müller, and Lenz 2023).

While previous research has examined specific manifestations like hate speech (Basile et al. 2019; Waseem and Hovy 2016) and fear speech (Saha et al. 2021, 2023), these represent components of the broader process of *othering*—the systematic construction of an outgroup as fundamentally different and threatening to the ingroup. Historically, othering was used to justify violence and repression in various societies. In Nazi Germany, discursive strategies mobilized hatred against Jews, Romani and homosexuals (Reicher, Haslam, and Rath 2008), while Stalinist terror was framed as a moral battle against "wreckers, diversionists and spies" (Gerwarth 2007). In India, Hindutva organizations (Hindu nationalist ideology) invoked narratives of Hindu tolerance to justify violence against Muslims, blaming conflicts on supposed Muslim intolerance (Kaur 2005).

In recent years, social networks have been used to spread narratives depicting immigrants and minorities as threats to national values (Pettersson and Sakki 2017), with extreme consequences such as incitement to violence against the Rohingya minority in Myanmar (Yue 2019). We ground our analysis in sociological theory, defining othering as the process of depicting a group as fundamentally different from one's own (Reicher, Haslam, and Rath 2008; Jetten, Spears, and Manstead 1997), characterized by positive ingroup and negative outgroup portrayals. This process marginalizes groups based on arbitrary differences like race, religion, or ethnicity, reinforcing social hierarchies (Duckitt 2003). Although existing research has explored explicit manifestations of intergroup conflict, our work introduces a taxonomy of othering language that extends beyond traditional indicators.

Additionally, we develop a computational framework that leverages large language models (LLMs) as classifiers, which facilitates rapid adaptation to new domains. After validating the model, we use it to explore the language of intergroup conflict in real-world scenarios. We analyze a corpus of messages posted on Telegram by Russian and Ukrainian war bloggers during the ongoing war between Russia and Ukraine, as well as a corpus of messages posted on the social media platform Gab. We explore the following research questions:

1. How does the use of othering language by Russian and Ukrainian war bloggers on Telegram change over the course of the war?

2. How does the moral and othering language used by war bloggers interact and vary by group?
3. How does portraying the target group as the 'other affect social attention?
4. Does use of othering language intensify during times of crisis, and in what ways are these behaviors more strongly rewarded?

Our analysis reveals the amplification of othering language in polarized online environments and its tendency to attract attention, especially during crises. While we find that othering language is often moralized across groups, its asymmetrical use by Russian war bloggers highlights its distinct utility in propaganda. By exploring these dimensions, we demonstrate how othering language grounds more overt expressions like fear speech, hate speech, and exclusionary practices, offering crucial insights for developing strategies to counteract othering and mitigate its impact on social cohesion.

Related Work and Background

The Language of Intergroup Conflict Intergroup conflict often drives violence by framing outgroups as existential threats to the ingroup. From Nazi Germany to modern online spaces, fear-mongering and hateful rhetoric radicalize populations by portraying outgroups as dangerous and immoral, justifying hostility and violence (Greipl, Rothut, and Schulze 2022; Reicher, Haslam, and Rath 2008).

Such conflict escalates during crises—pandemics, financial collapses, or political upheaval—when fears are externalized and outgroups are scapegoated for societal problems. Economic hardship heightens competition and prejudice, as seen during the Great Depression. Similarly, the 2015 European Refugee Crisis saw the rise of exclusionary rhetoric, driven by political and social tensions (Pettersson and Sakki 2017). Misperceptions of outgroup hostility further exacerbate tensions, with groups mistakenly believing the other supports violence, as seen in the 2021 Israeli-Palestinian conflict. However, corrective interventions have shown promise in reducing these tensions (Nir et al. 2023).

Outgroup threat perception—the belief that outgroups endanger the ingroup's identity or existence—serves as a key psychological mechanism driving this conflict. This dynamic was particularly evident during the COVID-19 pandemic, where heightened awareness of mortality intensified xenophobia (Esses and Hamilton 2021). These manufactured perceptions of threat can escalate conflict, leading to scapegoating that legitimizes violence and perpetuates hatred, resulting in real-world violence that reinforces instability and deepens social divisions (Fink 2018; Warofka 2018).

Computational scientists often operationalize the language of intergroup conflict through hate speech and fear speech. Hate speech refers to expressions intended to insult, degrade, or incite hostility toward groups based on attributes such as race, religion, gender identity, sexual orientation, disability status, or other social categories (Mathew et al. 2021). It promotes explicit hostility through vilification and dehumanization (Basile et al. 2019; Waseem and Hovy 2016; Kennedy et al. 2018). Fear speech invokes existential fear, portraying the target group as a fundamental threat to the ingroup's survival, culture, or identity (Buyse 2014; Saha et al. 2021, 2023). Both forms reinforce an 'us-versus-them' mentality by inciting hostility or instilling fear, emphasizing the perceived danger to the ingroup's way of life. Together, they contribute to othering, where groups are marginalized based on perceived differences (Reicher, Haslam, and Rath 2008; Jetten, Spears, and Manstead 1997).

Social Mechanisms of Othering Othering language encompasses the broader sociological process of constructing and excluding outgroups based on various social identities and characteristics. This process has been observed throughout history, from the Holocaust to contemporary political discourse on immigration (Reicher, Haslam, and Rath 2008). Understanding these mechanisms is essential for mitigating their effects and preventing intergroup conflict.

We base our understanding of othering on Reicher's model (Reicher, Haslam, and Rath 2008), which conceptualizes hate as emerging from a continuous process. As illustrated in Figure 1, extreme violence and genocide are driven by a distorted perception of group identity, where individuals are targeted solely for their perceived group membership (Reicher, Haslam, and Rath 2008; Kaur 2005). In this process, even disavowing one's group identity offers no protection, as othering reduces individuals to their perceived group membership, overriding personal actions or beliefs (Reicher, Haslam, and Rath 2008; Duckitt 2003).

The model outlines five key steps in the development of collective hate: (i) creating a cohesive ingroup, (ii) excluding specific populations, (iii) framing the outgroup as a threat to the ingroup's existence, (iv) portraying the ingroup as virtuous, and (v) celebrating the outgroup's destruction as a defense of ingroup virtue (Reicher, Haslam, and Rath 2008). Central to this model is the formation of both ingroup and outgroup boundaries, which together foster hostility and threat perception (Reicher, Haslam, and Rath 2008). This dynamic manifests in various ideologies where outgroups are framed as existential threats, justifying violence as moral and necessary for protecting the ingroup's values (Kennedy et al. 2023). Contemporary threats—whether economic or cultural—are often constructed to legitimize hostility. These identity narratives actively adapt to the ingroup's fears (Reicher, Haslam, and Rath 2008).

The intersection of othering and moralized language is crucial for understanding how narratives of intergroup conflict are constructed. Moral language often serves as a powerful tool for legitimizing exclusion and violence (Fiske and Rai 2014; Kennedy et al. 2023). According to the *Moralized Threat Hypothesis*, extreme expressions of prejudice are frequently driven by the belief that the outgroup has committed a moral transgression, making violence not only justified but also morally righteous (Hoover et al. 2021). Moreover, othering's role in capturing social attention demands further study. Recent computational work has made significant progress in detecting and analyzing othering language online through threat-based embeddings (Alorainy et al. 2019) and cross-platform analysis of cyber hate (Burnap and Williams 2016). Other studies show how narratives

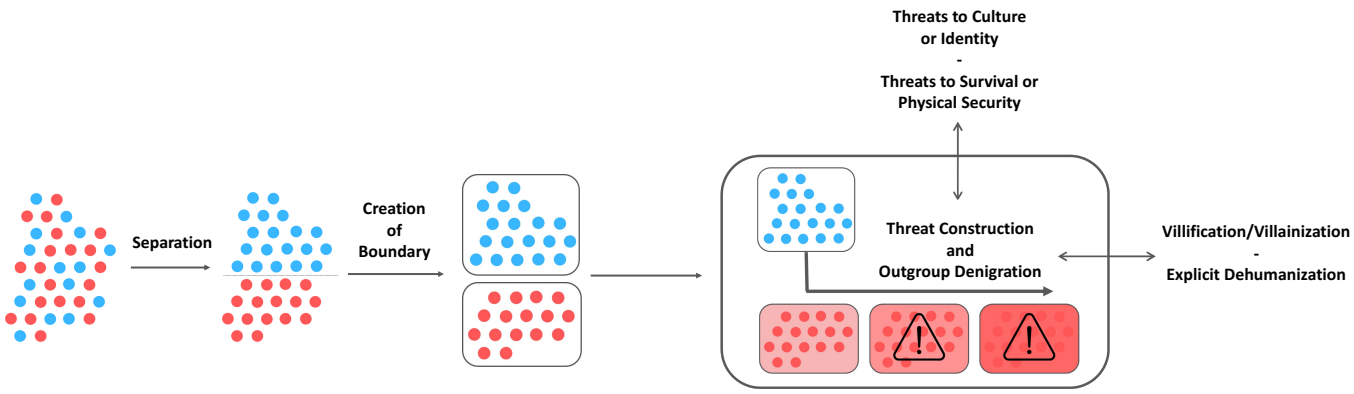


Figure 1: Conceptualization of the othering process. Othering starts with the separation of ingroup and outgroup members, the creation of symbolic boundaries, and the subsequent pipeline of othering. Through the construction and affirmation of perceived threats, the outgroup is increasingly framed as a threat.

built around othering often attract disproportionate attention in online spaces (Saha et al. 2023). This heightened attention amplifies their spread and impact, particularly during times of crisis or polarization. Understanding this dynamic is essential for grasping how othering influences public discourse and fosters division.

LLMs and In-Context Learning In-context learning enables LLMs to perform tasks by leveraging a few input-label pairs, or demonstrations, without requiring gradient updates, and has often outperformed zero-shot learning across numerous tasks (Zhao et al. 2021). However, its success is influenced by factors such as task complexity, the quality of provided examples, and the number of demonstrations. Our approach, which uses the *system prompt* to guide an LLM fine-tuned on one domain to adapt to another, offers a similar but distinct method for applying LLMs to unseen and untrained data.

Methods

Data

Russia-Ukraine war bloggers We analyze posts from Russian-oriented and Ukrainian-oriented Telegram channels (Theisen et al. 2022) from October 2015 to August 2023, comprising 989 channels and 9.67 million posts primarily in Ukrainian and Russian. This dataset is publicly released with this paper. Telegram has become crucial for military bloggers and news consumption in Russia and Ukraine, especially after bans on other platforms like Facebook and TikTok since 2020 (Oleinik 2024). We constructed a network of channels where directed links with weight w connect channel A to B if A forwards B's posts w times. The resulting network (Figure 9) reveals two main clusters.

Domain experts manually labeled 100 random channels as "Pro-Russia," "Pro-Ukraine," or "Other" based on channel bios and recent posts, focusing on explicit support or opposition to each side's narratives. Using these as seeds, we employed label propagation (Garza and Schaeffer 2019) to classify the remaining channels, achieving over 90% agreement when validated against expert reviews of 200 additional randomly selected channels. The final dataset contains 243 pro-Ukrainian channels (4.2M posts) and 325 pro-Russian channels (4.4M posts), with most activity concentrated around and after the 2022 Russian invasion. While messages prior to late 2021 are relatively sparse, we include them for completeness, recognizing that some bloggers transitioned to war-related coverage only after the outbreak of hostilities.

Gab Corpus Introduced in prior work by Saha et al. (Saha et al. 2023), this corpus contains 9,441 text posts from the popular 'alt-tech' social media platform Gab (Dehghan and Nagappa 2022). Gab, which is popular with political conservatives in US, hosts discussions often revolving around issues of race, immigration, and national identity. Despite the absence of direct physical conflict, the rhetoric on Gab is steeped in othering narratives, making it a platform of choice for studying hate speech and fear speech (Kennedy et al. 2018; Saha et al. 2023). Each post was manually classified by annotators into one or more categories: 'normal,' 'fear speech,' or 'hate speech,' with some posts receiving multiple labels. We detail the definitions used for fear speech and hate speech in the section **Relationship to Language of Intergroup Conflict**. Overall, 44.8% of the posts are labeled as normal, 19.7% as fear speech, and 42.4% as hate speech

A Model of Othering

We develop a flexible, LLM-based model to recognize othering language in text.

Taxonomy of Othering Othering is a group self-talk process that helps shape group's conceptualization of itself as good and virtuous and the other group as inherently evil and dangerous. When talking about itself, i.e., the ingroup, the group uses fear-laden speech (Lerman et al. 2024), which serves to bind the group together, often in response to a perceived threat from the outgroup. When talking about the

other, i.e., the outgroup, othering manifests itself through animosity and hostility (Stephan, Ybarra, and Rios 2015; Joffe 1999). To capture the various dimensions of othering, we define four categories of language linked to the othering process and provide translated examples from Russian war bloggers, and we provide examples of each class from the data in Table 6. The first two categories address perceived threats to the ingroup, while the latter two focus on the demonization of outgroups.

Threats to Culture or Identity arise when the outgroup is framed as a danger to the ingroup's cultural or social survival, challenging its values, language or traditions (Wohl, Branscombe, and Reysen 2010; Reicher, Haslam, and Rath 2008; Stephan, Ybarra, and Rios 2015; Joffe 1999).

Threats to Survival or Physical Security involve portraying the outgroup as an existential menace to the ingroup's physical well-being, thereby justifying preemptive hostility (Wohl, Branscombe, and Reysen 2010; Reicher, Haslam, and Rath 2008; Stephan, Ybarra, and Rios 2015; Joffe 1999).

Vilification/Villainization casts the outgroup as inherently evil or immoral, which in turn validates resistance and aggression (Joffe 1999; Reicher, Haslam, and Rath 2008; Stephan, Ybarra, and Rios 2015).

Explicit Dehumanization represents the most extreme form of othering, where the outgroup is compared to animals, objects or spirits, paving the way for extreme violence (Joffe 1999; Reicher, Haslam, and Rath 2008; Stephan, Ybarra, and Rios 2015).

These classes are integral to understanding how outgroups are systematically constructed, legitimizing their exclusion and violence.

Artificial Annotator Alignment Process We train an LLM classifier to recognize othering language using an "Artificial Annotator Alignment" process, inspired by knowledge distillation (Gou et al. 2021). We chose to use an LLM over more traditional NLP models for two key reasons. First, our evaluation of traditional models revealed that they struggled to capture the nuanced language underlying othering, as demonstrated in Tables 17 and 18, where XLM-RoBERTa (Conneau et al. 2020) performed significantly worse in the classification task compared to LLMs. Second, we sought to leverage the distinctive transfer learning capabilities of LLMs.

As illustrated in Figure 10, this process 'trains' LLMs as annotators by requiring them to align with human-labeled data. First, human annotators label a small subset of the dataset. We then validate a "high-quality" LLM (HQ-LLM), such as GPT-4, on that subset; if its annotations closely match human labels, it is used to annotate a larger dataset. Finally, we fine-tune an open-source LLM (OS-LLM), for example Llama3, on these expanded annotations. By combining HQ- and OS-LLMs, we scale up from limited human labels to a larger annotated corpus with minimal manual cost. Our ultimate goal is to deploy a cost-effective OS-LLM that matches the HQ-LLM's annotation quality and remains faithful to human judgment.

This approach is guided by two core principles: (1) The initial, high-quality (HQ) LLM must produce annotations that closely resemble those made by humans; (2) The open-source (OS) LLM fine-tuned on the HQ-LLM-annotated data must achieve performance on par with the HQ-LLM when evaluated against the human-annotated data (which is held-out throughout the process for evaluation). To adhere to the first principle, we assess the continuity between human annotations and the HQ-LLM using both standard machine learning metrics (accuracy, F1-score, etc.) and inter-annotator agreement metrics. This approach allows us to evaluate the HQ-LLM both as a classifier and as an artificial annotator, ensuring its consistency with human annotations. If both sets of evaluations yield optimal results, we consider the HQ-LLM a reliable proxy for human annotation. We then proceed to use the HQ-LLM to annotate a substantially larger portion of the dataset.

To adhere to the second principle, after fine-tuning the target OS-LLM on the data annotated by the HQ-LLM, we test its performance on the human-annotated dataset, once again using both sets of metrics to ensure that its classification metrics do not degrade from the HQ-LLM (ensuring inter-model continuity) and that its classifications are consistent with human annotations (ensuring inter-annotator continuity). This two-step validation ensures that the target model OS-LLM not only replicates the quality of the HQ-LLM's annotations but also aligns with human judgment, thereby confirming its effectiveness in the domain.

Data Annotation Our analysis used datasets of Russian and Ukrainian war blogger messages, combining random sampling with keyword-based upsampling. Keywords included coded denigration terms (e.g., "Ukronazis" in Russian data), identified through literature and domain expert consultation. To ensure data quality despite potential keyword-based false positives, we conducted thorough human annotation. Six annotators labeled 316 Russian posts and three annotators labeled the Ukrainian posts, with overlapping assignments. Inter-annotator agreement metrics (Tables 9, 12) aligned with similar studies (Saha et al. 2021, 2023). Final labels were determined by majority vote (Tables 7, 11), and show a balanced dataset validated through human verification.

Category	Instance Counts
Threats to Culture or Identity	45
Threats to Survival or Physical Security	41
Vilification/Villainization	52
Explicit Dehumanization	32
None	87
Total Data Points	**212**

Table 1: Human-annotated gold set summary for Ukrainan war bloggers data.

High-Quality LLM Annotation: Following independent human annotations, we used GPT-4o (HQ-LLM) to anno-

tate the same examples using prompts tailored to the specific context (prompts available in our GitHub repository), outputting a dictionary for each post : "'Threats to Culture or Identity': 1, 'Threats to Survival or Physical Security': 0, 'Vilification/Villainization': 1, 'Explicit Dehumanization': 0, 'None': 0", along with an explanation. For example, "The text describes local Nazis desecrating a historic Russian cemetery, in a way that represents a threat to cultural identity and vilifies the opposing group." These explanations were crucial for understanding the rationale behind each annotation and for testing our Rapid Domain Adaptation method later. The annotations were validated using metrics, such as Cohen's Kappa, treating GPT-4o as an additional annotator. The results, detailed in tables 13, show that the GPT-4o annotations were reliable and consistent across domains. In total, GPT-4o annotated approximately 20,000 posts (10,000 from each dataset) at a cost of approximately $70 USD, significantly lower than human annotation costs while remaining consistent with human annotators.

Fine-Tuning Models We fine-tuned three OS-LLMs (Mistral[1], LLaMA3-8b-Instruct[2], and LLaMA2[3]) on the GPT-4o-annotated data using LoRA. An example instruction-answer pair is shown in Example 2. Each model was fine-tuned on three datasets: Russian-only, Ukrainian-only, and combined Russian-and-Ukrainian (detailed in Tables 14, 15, and 16). This setup enabled testing domain generalization and developing our Rapid Domain Adaptation methodology. Common test and validation sets were withheld to ensure fair evaluation. Models were fine-tuned for 5 epochs on consumer hardware (NVIDIA RTX 3090, though viable on 16GB VRAM GPUs) using a 0.7:0.1:0.2 split for fine-tuning, validation, and testing. LoRA parameters included adapter rank 8, 4-bit quantization, and settings from Table 8, maintaining performance within 1-2

LLaMA3-8b-Instruct performed best (Table 2), showing minimal degradation compared to GPT-4o. Domain-specific models excelled in their domains but struggled with cross-domain generalization, while multi-domain models showed better generalization at the cost of domain-specific performance. Our LoRA implementation enables reproducibility with modest computational resources.

Category	Cohen's	Accuracy	F1 Score
Threats to Culture or Identity	0.84	0.89	0.86
Threats to Survival or Physical Security	0.74	0.84	0.80
Vilification/Villainization	0.81	0.91	0.89
Explicit Dehumanization	0.84	0.89	0.88
None	0.84	0.90	0.87

Table 2: Inter-Annotator Agreement and Model Performance: Cohen's Kappa (Agreement), Accuracy, and F1 Score between Majority Vote and LLM on Russian Data. For F1 and Accuracy, five trials were ran, and with a temperature set to 0, identical results were obtained in every trial.

[1]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[2]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[3]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

Category	Cohen's	Accuracy	F1 Score
Threats to Culture or Identity	0.81	0.88	0.89
Threats to Survival or Physical Security	0.72	0.87	0.82
Vilification/Villainization	0.80	0.87	0.86
Explicit Dehumanization	0.82	0.89	0.88
None	0.83	0.88	0.84

Table 3: Inter-Annotator Agreement and Model Performance: Cohen's Kappa (Agreement), Accuracy, and F1 Score between Majority Vote and LLM on Ukrainian Data. For F1 and Accuracy, five trials were ran, and with a temperature set to 0, identical results were obtained in every trial.

Rapid Domain Adaptation Adapting models to new domains presents a significant challenge in classification tasks. To study othering across multiple domains in a cost-effective manner, we test whether our models could effectively transfer knowledge from one domain to another. LLMs are well-suited for this task due to their (1) inherent complexity and (2) the pseudo-world mapping generated through extensive pretraining on vast corpora. To leverage this power, we employ two techniques that adapt the classifier to new, unseen data: system prompt steering and logit disambiguation. We name this approach Rapid Domain Adaptation (RDA).

System Prompt Steering: The first component of our RDA system involves manipulating the system prompt for the fine-tuned model. A system prompt provides the model with initial instructions, guiding how it processes and classifies incoming data. While in-context learning offers some benefits, simply appending context to new data only minimally improves performance. However, a well-crafted system prompt (we illustrate this in Example 1 in the Appendix), designed to steer the model's reasoning, led to significant improvements in new domains by explicitly framing the new domain (this logic is illustrated in Figure 12) in relation to the model's prior training, we achieved notable enhancements in domain adaptation.

Logit Disambiguation: The second component of our RDA system involves exposing the logits for each class, rather than simply assigning a binary label of 0 or 1. Logits represent the model's confidence in its predictions, indicating the likelihood of a particular token being chosen. Since our task is to classify messages using either 1 or 0 for each class, we can expose the logits for this classification token and then using confidence thresholds for each class, fine-tuning them specifically for new domains. This approach is especially useful when the new domain features significantly different language or topics compared to the original training (fine-tuning) domain. As shown in Figure 11, by disambiguating the logits and adjusting confidence thresholds, we can better adapt the model to the new domain. Overall, these components — system prompt steering and logit disambiguation — work together to enable rapid and reliable domain adaptation, leveraging modern LLMs to effectively handle drastic shifts in domain context.

RDA Evaluation: We evaluated our RDA system across all cross-domain combinations (e.g., a model fine-tuned on Russian war blogger data performing on Ukrainian war blogger data), with detailed results available in our repos-

itory. We first compared system prompt steering to two baseline approaches: no added context and traditional in-context learning, where context is simply appended to the new prompting data. Table 21 shows the F1 scores for different domain transitions. Figure 2 visualizes the substantial performance gains across metrics when a model fine-tuned on Russian data applied to Gab, which contains messages posted on a different platform, in a different language (English), and in a different cultural and geopolitical context (representing the most substantial domain transition). These results demonstrate the effectiveness of our RDA system, especially in models that had no prior exposure to the new domain.

Prompting Type	Accuracy	F1 Score	Precision	Recall
No Additional Prompt	0.55	0.53	0.83	0.56
In-Context Learning	0.61	0.63	0.84	0.65
System Prompt	0.69	0.76	**0.90**	0.70
RDA	**0.71**	**0.77**	**0.90**	**0.71**

Table 4: Performance Comparison of Different Prompting Types: Accuracy, F1 Score, Precision, and Recall. For metric scores, five trials were ran, and with a temperature set to 0, identical results were obtained in every trial.

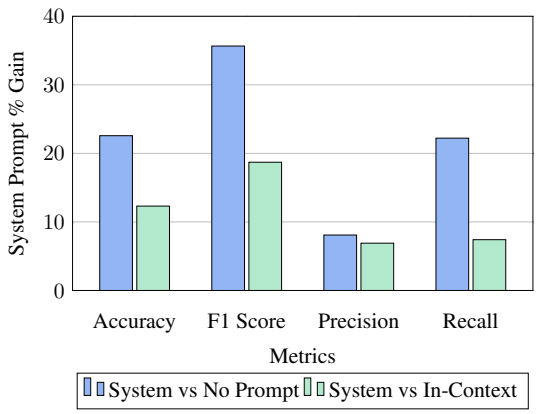


Figure 2: System Prompt % Gain Compared to No Additional Prompt and In-Context Learning Across Metrics (Accuracy, F1 Score, Precision, Recall).

Relationship to Language of Intergroup Conflict

We examine the relationship between othering language and other widely studied expressions of intergroup conflict, such as fear speech and hate speech, using the annotated Gab corpus. As illustrated in Fig. 3, fear speech and hate speech reflect expressions of othering, but they do not fully encompass it. Instead, they function as partial components of the broader process. This underscores the asymmetry suggested by our model, which posits that othering language subsumes, but is not limited to, specific expressions like fear speech and hate speech. Additionally, we consider the practical implications for content moderation by evaluating how well current toxicity classifiers detect othering language

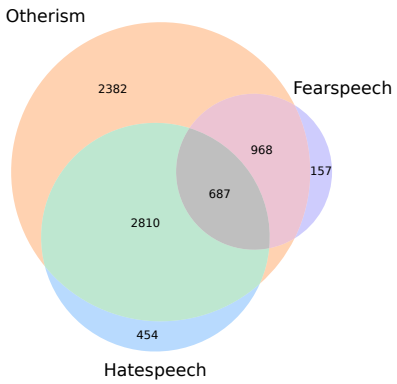


Figure 3: Venn diagram showing overlap between othering, fear speech, and hate speech in the Gab corpus. The diagram reveals that while fear speech and hate speech often co-occur with othering, many instances of othering occur without these explicit forms of conflict language.

alongside fear speech and hate speech. This analysis highlights the limitations of existing classification tools in addressing the full spectrum of othering language.

Fear Speech Fear speech, defined as expressions that instill existential fear of a target group (often based on attributes such as race, religion, or gender) (Buyse 2014), is significantly associated with othering language. Specifically, the probability that fear speech contains 'Vilification/Villainization' is 68.9%, and the probability of containing 'Threats to Culture or Identity' is 50.5%. It has a weaker association with 'Threats to Survival or Physical Security' at 20.3%. These connections reflect the nature of fear speech in both evoking existential fear and subtly vilifying the outgroup, such as in messages like "They will destroy our way of life unless we stop them." Moreover, fear speech shows an asymmetric relationship to othering: 88.9% of fear speech instances involve othering, but only 24.2% of othering messages are classified as fear speech.

Hate Speech Hate speech is language used to express hatred toward a targeted individual or group or is intended to be derogatory, to humiliate, or to insult the members of the group, based on attributes such as race, religion, or gender (Mathew et al. 2021), and is typically the most explicit form of othering. Our analysis shows strong associations with 'Vilification/Villainization' (74.1%), 'Explicit Dehumanization' (37.3%), and 'Threats to Culture or Identity' (32.3%). These connections underscore hate speech's dual role both denigrating the outgroup and framing it as a threat.

Hate speech also demonstrates an asymmetric relationship with othering language: approximately 87.4% of hate speech involves othering, but only 51.1% of messages containing othering language are classified as hate speech.

Toxicity Finally, we analyze the relationship to toxicity. Using the Detoxify classifier[4], which rates text on a scale

[4]https://github.com/unitaryai/detoxify

from 0 to 1 (with scores above 0.5 considered toxic), we find the following average toxicity scores: fear speech averages 0.46, while hate speech scores higher at 0.65. For othering content, excluding Explicit Dehumanization, the average toxicity is 0.42 and increases to 0.53 when Explicit Dehumanization is included (which has a notably high average toxicity of 0.81). These findings highlight the broad spectrum of toxicity within othering rhetoric and emphasize the need for more effective detection tools.

Results

We label the corpus of messages posted on Telegram by Russian and Ukrainian war bloggers for othering language, focusing on the period from late 2021 to August of 2023. This period of war was characterized by high conflict and animosity between the groups. We also label a corpus of messages posted on Gab, which is often favored by far-right users within the US.

Othering During the Russia-Ukraine War

Figure 4 shows that othering rhetoric fluctuated throughout the conflict, rising after key events. We define key events as significant events during the war that also became prominent discussion topics within their respective communities (i.e., Russian and Ukrainian war bloggers separately). Here, we utilize the events enumerated in Tables 23 and 24, which were compiled using domain knowledge in conjunction with methods from previous work (Gerard et al. 2024).

The types of events that tend to correlate with spikes in othering rhetoric differ significantly between these communities. For Russian bloggers, the most prominent driver was US military and security aid to Ukraine, which often led to a rise in rhetoric, as they framed themselves as victims of Western aggression. In contrast, for Ukrainian bloggers, increases in othering language were primarily driven by military gains and losses on the battlefield, reflecting a more direct response to the dynamics of the war itself. These findings are consistent with previous research showing that Russian bloggers tend to react more to international actions, while Ukrainian bloggers are more focused on military developments (Gerard et al. 2024). In future work, we plan to automate event detection using change point analysis and further explore the causal relationship between these events and the prevalence of othering language.

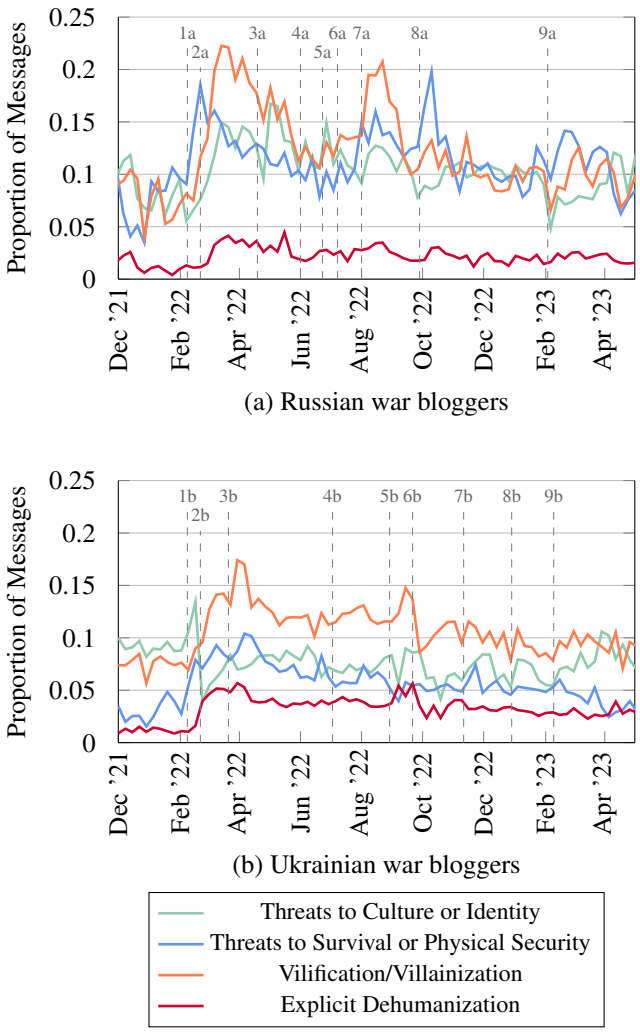


Figure 4: Temporal trends in the proportion of messages with othering language posted by (a) **Russian war bloggers** and (b) **Ukrainian war bloggers** from December 2021 to May 2023. The four classes of othering language are: Threats to Culture or Identity, Threats to Survival or Physical Security, Vilification/Villainization, and Explicit Dehumanization.

The Moral Language of Othering

We explore the interaction between othering and moral language on Telegram and Gab. For both platforms, we label moral values expressed in text using a model fine-tuned to recognize moral language (Trager et al. 2022). The model identifies the moral foundations of human intuitive ethics, such as valuing of purity, respect for authority, equality (fairness), group loyalty, and care (Graham et al. 2013).

Moralized Othering on Telegram We begin by analyzing messages posted by war bloggers on Telegram, using a chi-squared test to explore the relationship between moral language in messages containing othering language versus those without it. The chi-squared test helps determine whether the association between moral language and othering is statistically significant, rather than occurring by chance. The results indicate a strong, statistically significant connection ($p < 0.001$), demonstrating that moral language is more likely to co-occur with othering language than with non-othering language. This suggests that moral framing is frequently used to justify or intensify othering language, supporting the Moralized Threat Hypothesis (Hoover et al. 2021).

Next, we analyzed the use of moral language by Russian and Ukrainian war bloggers across different othering categories to identify differences in their moral framing strategies. Figure 6 highlights the significant variations in the

moral values expressed by each side when using othering language, based on log-odds ratios. We also examined the interaction between specific moral language categories and individual othering classes. As shown in Figure 14, the two groups also differ significantly in their use of moral language within othering rhetoric, with these differences confirmed by two-proportion z-tests ($p < 0.001$).

Explicit forms of othering, such as *Explicit Dehumanization*, are the least associated with moral language in both groups, likely due to their overtly aggressive nature, which doesn't align well with broader moral frameworks (Saha et al. 2023). However, for Russian war bloggers, the strongest association between morality and othering is found in the purity moral frame within *Explicit Dehumanization*. This suggests that while Russians use moral language less frequently with dehumanization, when they do, it is often tied to purity, reflecting popular narratives portraying Ukrainians (and the West) as 'puppets' or 'zombies' corrupting Russian values. Meanwhile, for Ukrainian war bloggers, the most morally charged category is *Threats to Survival or Physical Safety*, most closely associated with care, which aligns with the context of Russia's invasion.

Overall, both groups display similar trends, but the use of moral language reveals strategic differences. Russian bloggers emphasize purity and cultural threats, reinforcing existential threat and victimization (Geissler et al. 2023), while Ukrainian bloggers focus on care in response to physical threats. Our analysis shows that moral language and othering are deeply intertwined, with Russian bloggers heavily relying on moralized language to justify intergroup prejudice. This suggests moralized othering is an effective propaganda tool, though further research is needed to understand its role in polarized environments.

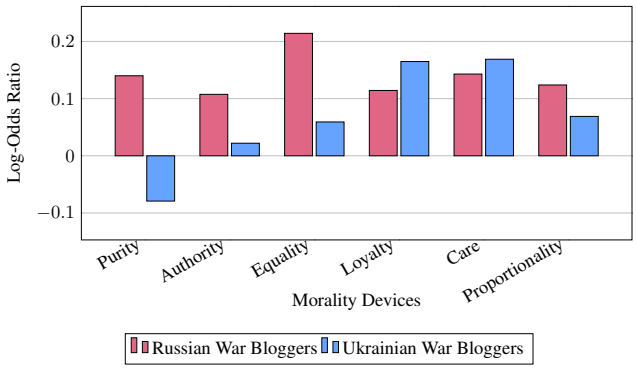


Figure 5: Log-odds ratios of morality devices for Russian and Ukrainian war bloggers, comparing the presence of moral language in messages with othering language versus those without. Ukrainian war bloggers are represented in light blue, and Russian war bloggers in light red.

Moralized Othering on Gab Gab provides an additional lens for examining the interplay between moral language and othering in a polarized environment with lower levels of immediate conflict. Although not directly involved in intergroup violence, the platform's rhetoric is steeped in other-

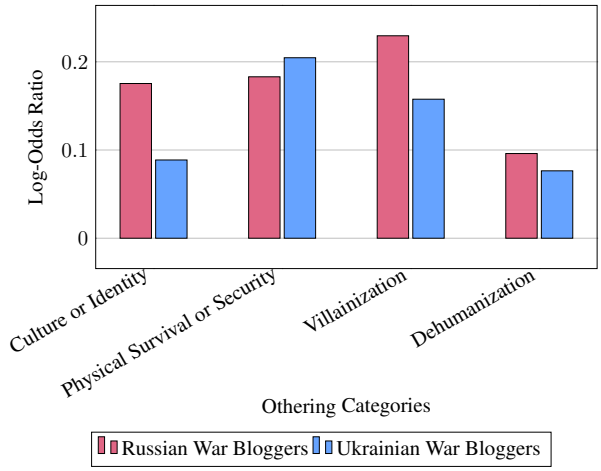


Figure 6: Log-odds ratios of morality use for Russian and Ukrainian war bloggers, comparing its use in messages with specific othering language use versus those without. Ukrainian war bloggers are represented in light blue, and Russian war bloggers are represented in light red.

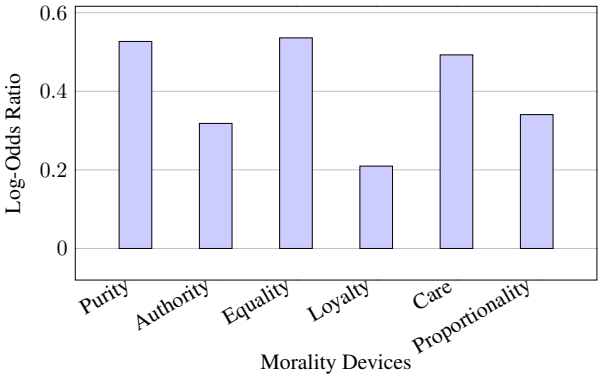


Figure 7: Log-odds ratios of morality devices for Gab, comparing the presence of moral language in messages with othering language versus those without.

ing narratives within an American context. This makes Gab an ideal setting to further test our hypothesis in a context similarly shaped by divisive and exclusionary discourse, but without direct conflict.

As in the previous case study, we first applied the chi-squared test to confirm a significant difference in the use of moral language between messages with and without othering language ($p < 0.001$). We then evaluated the log-odds ratios of moral language in othering messages and the interaction between specific moral categories and othering types, as shown in Figures 7 and 15. The most morally loaded othering category is *Threats to Survival and Physical Security*, primarily associated with care, similar to Ukrainians. *Threats to Culture or Identity* is linked to equality and loyalty, similar to Russians. *Explicit Dehumanization* is tied to purity but not to other moral values, unlike in Russians, where it had (weak) links to all moral values.

These results not only support the hypothesis that moral language and othering are closely intertwined but also reaffirm the findings from the previous case study, highlighting how moral language shapes and sustains othering narratives. They also offer a new lens on how different moral devices are strategically used to appeal to different audiences, underscoring the need for further research into how these narratives are constructed and spread in polarized environments.

Othering and Attention

Next, we explore the relationship between online attention and the use of othering language in messages by Russian and Ukrainian war bloggers.

Network Centrality We measure channel centrality in the reference network of war blogger channels (see Methods) using degree centrality and eigenvector centrality. (For simplicity, we use an undirected version of this network.) Degree centrality reflects the influence distribution and communicative ability of nodes in the network and eigenvector centrality captures the positional importance of network nodes. We then calculate the Spearman correlation between channel centrality and its use of othering language (as a proportion of its messages).

Network Centrality The results, shown in Table 5, reveal statistically significant correlations between both degree centrality and eigenvector centrality and the use of othering language by war bloggers. Notably, the correlation is stronger among Russian war bloggers. This suggests that those who employ othering language may occupy more influential positions within the Telegram network, but the direction of this relationship remains unclear. It could be that established opinion leaders are more inclined to use inflammatory othering language, or that this language itself drives increased attention and thus leads to higher centrality. Our current design cannot disentangle these possibilities, and further longitudinal or experimental work would be needed to pinpoint the causal direction.

Message Views To further investigate the link between othering and attention, we focus on message-level metrics. Each message's number of views is normalized by the channel's typical viewership using Z-score normalization. As shown in Figure 8, messages containing othering language consistently garner more views than those without, a relationship confirmed by the Mann-Whitney U Test ($p < 0.001$). However, while these results indicate that messages containing othering language receive greater attention, it remains unclear whether higher engagement stems from these messages' content, or whether users who already attract more attention are more inclined to post othering content. This study thus highlights a meaningful association but does not establish which factor causes the other.

Times of Crisis We define times of crisis as the week immediately following significant events during the war that became prominent discussion topics within the community (Russian and Ukrainian war bloggers separately). We identified key events by applying weekly topic clustering to track narrative evolution and used Discounted Cumulative Gain (DCG) to quantify engagement. Relevant events—major updates, turning points, or critical incidents—were manually filtered, excluding unrelated or trivial topics. The resulting events, listed in Tables 23 and 24, form the basis of our analysis. Overall, we observe that othering gains more attention immediately after a crisis.

As shown in Figure 22, degree centrality correlates more strongly with the proportion of othering messages following key events. Among Russian war bloggers, eigenvector centrality also rises significantly, whereas it slightly decreases for Ukrainian war bloggers. Additionally, Figure 13 shows that during heightened conflict, the viewership gap between messages with and without othering widens significantly for Ukrainian war bloggers (confirmed by a Mann-Whitney U Test), while remaining stable for Russian war bloggers.

This analysis indicates that othering not only correlates with increased attention but may also be structurally rewarded, especially during crises (although this would need to be further studied). Both network centrality and viewership metrics suggest that users employing othering language are more likely to gain influence and visibility. The sharp rise in attention, particularly among Ukrainian war bloggers during crises, underscores othering's role in driving engagement and influence in polarized environments.

	Centrality Metric	
Community	Degree	Eigenvector
Russian	0.254	0.333
Ukrainian	0.128	0.147

Table 5: Centrality and othering messages. Spearman correlation between a channel's proportion of messages with othering language and its degree and eigenvector centralities. All correlations are significant at the $p < 0.01$ level.

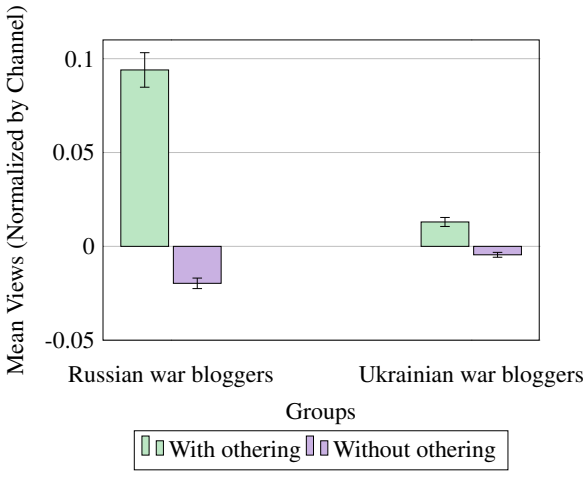


Figure 8: Comparison of mean views with and without othering (z-score channel-normalized). The bars represent the mean views for Russian and Ukrainian war bloggers, with and without othering, and the error bars indicate the standard error.

Discussion and Conclusion

Our work makes three distinct contributions to the field of intergroup conflict detection. First, we present a novel taxonomy of othering language that uniquely bridges sociological theory with computational methods, moving beyond traditional hate speech detection to capture subtle forms of outgroup discrimination. Second, we demonstrate through our computational framework how othering intersects with moral language and attention mechanisms in previously unquantified ways, revealing patterns that existing approaches have overlooked. Third, we introduce efficient methods for fine-tuning LLMs to detect and transfer knowledge about othering across domains, providing a scalable approach for studying this phenomenon in diverse contexts.

Our study identifies patterns of othering that align with well-documented phenomena—from social media's role in Myanmar's Rohingya crisis (Yue 2019) to polarization during the European refugee crisis (Pettersson and Sakki 2017). Although we focus on particular domains, the mechanisms we uncover—especially the escalation of othering during crises and its intersection with moral language—reflect broader theoretical frameworks of intergroup conflict (Reicher, Haslam, and Rath 2008). These parallels suggest that our findings may extend to other polarized contexts, though further research should validate this framework across diverse conflicts and domains.

In addition, our analysis underscores the fluidity of othering language and its responsiveness to real-world events. By demonstrating a consistent link between othering and moral language, we bolster the Moralized Threat Hypothesis and show how moral framing can intensify harmful narratives. Lastly, we observe a strong association between othering and social attention: such messages attract elevated engagement, particularly during crises when othering surges and gains disproportionate visibility.

Limitations

Our analysis focuses on Russian and Ukrainian war bloggers in a highly charged war environment, which may influence data neutrality. While the study covers the first year and a half of the conflict, the ongoing nature of the war means future developments could shift narrative dynamics. The centrality metrics we employ are derived from networks spanning the entire timeframe, which limits temporal granularity compared to platforms like Twitter, where interactions accumulate more rapidly. Due to Telegram's slower interaction patterns, smaller time slices did not yield dense enough temporal networks for robust analysis. Instead, we used alternative metrics to explore narrative and language evolution. Additionally, while we observe correlations between othering language and social attention (RQ3), we cannot determine whether othering language drives influence or if influential bloggers are more likely to employ it. Lastly, although our model allows for editing the system prompt in the RDA process, many open-source models (e.g., Mistral) lack this flexibility, potentially reducing adaptability in similar contexts. Biases may also appear, for example in our moral annotation process, where the model was trained on out-of-domain data; however, such biases are likely to cancel out in comparative analyses.

Summary and Future Work

This study advances computational social science through novel methods for detecting and analyzing online othering language. Our framework integrates sociological theory with machine learning to reveal new patterns: othering's heightened prevalence during crises, its intersection with moral rhetoric, and its role in attention mechanisms. These findings extend beyond traditional hate speech detection by capturing subtle forms of discrimination and their amplification through social media. Our methodological contributions—efficient LLM training and cross-domain knowledge transfer—provide tools for studying intergroup conflict across contexts. The framework has practical applications in conflict monitoring and intervention. It enables early detection of escalating tensions, informs platform moderation strategies through insights about moral framing and attention mechanisms, and provides scalable methods for adapting to new conflicts and cultural contexts. Future research should test these methods across different platforms and conflicts, explore intervention strategies leveraging moral-attention dynamics, and examine how group membership influences othering language—from core members' ideological rhetoric to boundary members' defensive language. Understanding these patterns could inform targeted interventions and reveal nuances in moralized othering online.

Acknowledgements

In memory of Kyrylo Serhiichuk (1990-2024).

References

Alorainy, W.; Burnap, P.; Liu, H.; and Williams, M. L. 2019. "The enemy among us" detecting cyber hate speech with threats-based othering language embeddings. *ACM Transactions on the Web (TWEB)*, 13(3): 1–26.

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Pardo, F. M. R.; Rosso, P.; and Sanguinetti, M. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, 54–63.

Burnap, P.; and Williams, M. L. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data science*, 5: 1–15.

Buyse, A. 2014. Words of violence:" Fear speech," or how violent conflict escalation relates to the freedom of expression. *Human Rights Quarterly*, 36(4): 779–797.

Cervone, C.; Augoustinos, M.; and Maass, A. 2021. The language of derogation and hate: Functions, consequences, and reappropriation. *Journal of language and social psychology*, 40(1): 80–101.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. arXiv:1911.02116.

Dehghan, E.; and Nagappa, A. 2022. Politicization and radicalization of discourses in the alt-tech ecosystem: A case study on Gab Social. *Social Media+ Society*, 8(3): 20563051221113075.

Duckitt, J. 2003. Prejudice and intergroup hostility.

Esses, V. M.; and Hamilton, L. K. 2021. Xenophobia and anti-immigrant attitudes in the time of COVID-19. *Group Processes & Intergroup Relations*, 24(2): 253–259.

Fink, C. 2018. Dangerous speech, anti-Muslim violence, and Facebook in Myanmar. *Journal of International Affairs*, 71(1.5): 43–52.

Fiske, A. P.; and Rai, T. S. 2014. *Virtuous violence: Hurting and killing to create, sustain, end, and honor social relationships*. Cambridge University Press.

Garza, S. E.; and Schaeffer, S. E. 2019. Community detection with the label propagation algorithm: a survey. *Physica A: Statistical Mechanics and its Applications*, 534.

Geissler, D.; Bär, D.; Pröllochs, N.; and Feuerriegel, S. 2023. Russian propaganda on social media during the 2022 invasion of Ukraine. *EPJ Data Science*, 12(1): 35.

Gerard, P.; Volkova, S.; Penafiel, L.; Lerman, K.; and Weninger, T. 2024. Modeling Information Narrative Detection and Evolution on Telegram during the Russia-Ukraine War. arXiv:2409.07684.

Gerwarth, R. 2007. The Dictators: Hitler's Germany and Stalin's Russia.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6): 1789–1819.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, 55–130. Elsevier.

Greipl, S.; Rothut, J.; and Schulze, M. 2022. Online Radicalization: The Role of Fear. *Journal of Online Behavior*, 10(3): 203–220.

Hoover, J.; Atari, M.; Mostafazadeh Davani, A.; Kennedy, B.; Portillo-Wightman, G.; Yeh, L.; and Dehghani, M. 2021. Investigating the role of group-based morality in extreme behavioral expressions of prejudice. *Nature Communications*, 12(1): 4585.

Jetten, J.; Spears, R.; and Manstead, A. S. 1997. Strength of identification and intergroup differentiation: The influence of group norms. *European journal of social psychology*, 27(5): 603–609.

Joffe, H. 1999. *Risk and'the Other'*. Cambridge University Press.

Kaur, R. 2005. *Performative politics and the cultures of Hinduism: Public uses of religion in western India*. Anthem Press.

Kennedy, B.; Atari, M.; Davani, A. M.; Yeh, L.; Omrani, A.; Kim, Y.; Coombs, K.; Havaldar, S.; Portillo-Wightman, G.; Gonzalez, E.; et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv. July*, 18.

Kennedy, B.; Golazizian, P.; Trager, J.; Atari, M.; Hoover, J.; Mostafazadeh Davani, A.; and Dehghani, M. 2023. The (moral) language of hate. *PNAS nexus*, 2(7): pgad210.

Lerman, K.; Feldman, D.; He, Z.; and Rao, A. 2024. Affective polarization and dynamics of information spread in online networks. *npj Complexity*, 1(1): 8.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14867–14875.

Nir, N.; Nassir, Y.; Hasson, Y.; and Halperin, E. 2023. Kill or be killed: Can correcting misperceptions of out-group hostility de-escalate a violent inter-group out-break? *European Journal of Social Psychology*, 53(5): 1004–1018.

Oleinik, A. 2024. Telegram channels covering Russia's invasion of Ukraine: a comparative analysis of large multilingual corpora. *Journal of Computational Social Science*, 1–24.

Pettersson, K.; and Sakki, I. 2017. Pray for the fatherland! Discursive and digital strategies at play in nationalist political blogging. *Qualitative Research in Psychology*, 14(3): 315–349.

Reicher, S.; Haslam, S. A.; and Rath, R. 2008. Making a virtue of evil: A five-step social identity model of the development of collective hate. *Social and Personality Psychology Compass*, 2(3): 1313–1344.

Saha, P.; Garimella, K.; Kalyan, N. K.; Pandey, S. K.; Meher, P. M.; Mathew, B.; and Mukherjee, A. 2023. On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, 120(11): e2212270120.

Saha, P.; Mathew, B.; Garimella, K.; and Mukherjee, A. 2021. "Short is the Road that Leads from Fear to Hate": Fear Speech in Indian WhatsApp Groups. In *TheWebConf 2021*, 1110–1121.

Schulze, M.; Müller, H.; and Lenz, S. 2023. Fear-Based Messaging in Extremist Communication. *Journal of Communication Studies*, 12(1): 99–115.

Stephan, W. G.; Ybarra, O.; and Rios, K. 2015. Intergroup threat theory. In *Handbook of prejudice, stereotyping, and discrimination*, 255–278. Psychology Press.

Theisen, W.; Cedre, D. G.; Carmichael, Z.; Moreira, D.; Weninger, T.; and Scheirer, W. 2022. Motif Mining: Finding and Summarizing Remixed Image Content. arXiv:2203.08327.

Trager, J.; Ziabari, A. S.; Davani, A. M.; Golazazian, P.; Karimi-Malekabadi, F.; Omrani, A.; Li, Z.; Kennedy, B.; Reimer, N. K.; Reyes, M.; Cheng, K.; Wei, M.; Merrifield, C.; Khosravi, A.; Alvarez, E.; and Dehghani, M. 2022. The Moral Foundations Reddit Corpus. arXiv:2208.05545.

Tsoy, D.; Tirasawasdichai, T.; Kurpayanidi, K. I.; et al. 2021. Role of social media in shaping public risk perception during COVID-19 pandemic: A theoretical review. *International Journal of Management Science and Business Administration*, 7(2): 35–41.

Warofka, A. 2018. An independent assessment of the human rights impact of Facebook in Myanmar. *Facebook Newsroom, November*, 5.

Waseem, Z.; and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, 88–93.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.

Wohl, M. J.; Branscombe, N. R.; and Reysen, S. 2010. Perceiving your group's future to be in jeopardy: Extinction threat induces collective angst and the desire to strengthen the ingroup. *Personality and Social Psychology Bulletin*, 36(7): 898–910.

Yue, N. 2019. The "Weaponization" of Facebook in Myanmar: A Case for Corporate Criminal Liability. *Hastings LJ*, 71: 813.

Zhao, Z.; Wallace, E.; Feng, S.; Klein, D.; and Singh, S. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, 12697–12706. PMLR.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Ethical Considerations
   (g) Did you discuss any potential misuse of your work? Yes, see Ethical Considerations
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see Ethical Considerations
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes, see Related Work and Background and Results
   (b) Have you provided justifications for all theoretical results? Yes, see Related Work and Background and Results
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, see Results
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, see Results
   (e) Did you address potential biases or limitations in your theoretical framework? Yes, see Limitations
   (f) Have you related your theoretical results to the existing literature in social science? Yes, see Related Work and Background and Results
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, see Results and Discussion and Conclusion
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, see Ethical Considerations

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see Methods

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see Methods

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see Methods

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see Methods

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? Yes, see Ethical Considerations

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see Ethical Considerations

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see Ethical Considerations

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? Yes, see Ethical Considerations

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? Yes, see repo

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Ethical Statement

Our study used publicly accessible data from influential figures discussing the Russia-Ukraine war on Telegram, following the FAIR data principles (Wilkinson et al. 2016). While we aim to advance computational analysis, we recognize the risks of misuse, such as evading classifiers or generating harmful language; applying our classifiers poses its own challenges: over-detection may suppress legitimate speech, while under-detection can spread harmful rhetoric. To ensure transparency and responsible use, we have made both the data and code publicly available, stressing the ethical application of these resources.

Appendix

Example 1: System prompt to steer model from Russo-Ukrainian context to Gab context.

You are a chatbot trained on how Russian warbloggers 'other' their enemy. Now you classify more subtle 'otherism' in United States users that are generally far-right white supremacists that dislike other races, liberals, refugees, muslims, gay people, etc. Additionally, you detect messages that aim to incite fear about these groups, often by emphasizing threats to security, culture, or values. Note whom they target in their messages and any elements that attempt to spread fear, uncertainty, or doubt about these groups.

Example 2: Example conversation showing hate speech annotation by the model.

User: "These Ukronazis tore apart our ancestors' resting place. They want to tear down everything dear to us. But when the Russians came, they shit themselves, just like their Bandera"

System: "{'Threats to Culture or Identity': 1, 'Threats to Survival or Physical Security': 0, 'Vilification/Villainization': 1, 'Explicit Dehumanization': 0, 'None': 0, 'explanation': 'The text describes local Nazis desecrating a historic Russian cemetery, in a way that represents a threat to cultural identity and vilifies the opposing group.'}"

Table 6: Example Posts for Each Category of Othering Language

Category	Example Post
Threats to Culture or Identity	"The erosion of the Russian language in Ukrainian schools: Ukrainian policymakers pushing to erase the Russian tongue risk severing the threads that weave together our history."
Threats to Survival or Physical Security	"Zelensky's regime has accumulated 30 tons of plutonium and 40 tons of enriched uranium at the Zaporizhia NPP [...] the regime really is on the verge of creating its own nuclear bomb! And hundreds of 'dirty' bombs can be made from such a quantity of radioactive material!"
Vilification/Villainization	"Because these Ukronazi girls can fight only by hiding behind hostages. All their courage went down the drain in chants and slogans like 'hang the Muscovite.' But when the Russians came, they shit themselves, just like their Bandera."
Explicit Dehumanization	"These are zombies, who may have been brothers before, but over the past 8 years, from the bite of Nazism and Banderization, they have turned into non-humans. That is why our army calls on all brothers to lay down their arms, so that we can distinguish a brother from an infected zombie, who can only bite and infect."

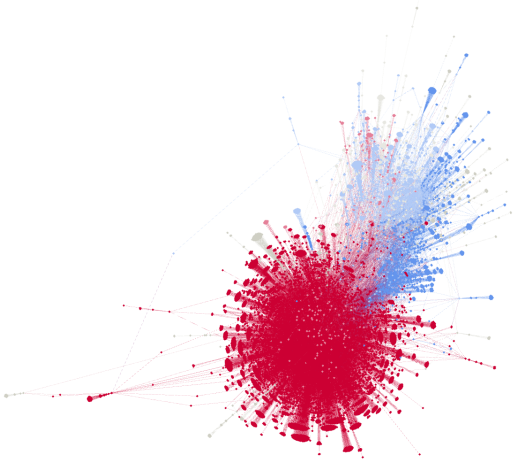


Figure 9: Visualization of a co-reference network based on message content and channel bios. Nodes are colored according to the stance indicated by their activity: Pro-Russian nodes are in red, Pro-Ukrainian nodes are in blue, and nodes not strongly affiliated to either nationality are in grey.Many of these grey, non-affiliated channels focus on day-to-day issues, such as local trading, equipment exchanges, or other non-political content. This network was constructed by analyzing messages and bio information from various Telegram channels.

Category	Instance Counts
Threats to Culture or Identity	122
Threats to Survival or Physical Security	62
Vilification/Villainization	164
Explicit Dehumanization	77
None	78
Total Data Points	**316**

Table 7: Summary of human annotations for Russian war bloggers data.

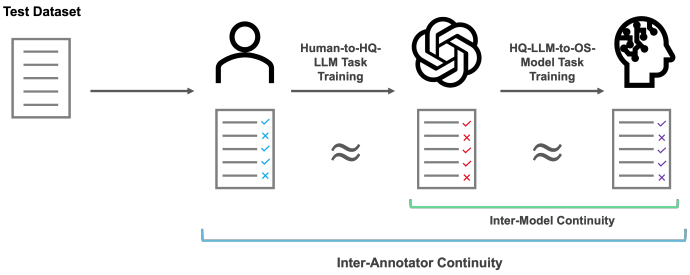


Figure 10: Artificial Annotator Alignment process. Human-annotated data is first used to fine-tune a high-quality LLM. The HQ-LLM's annotations are compared with human annotations for alignment, and then the HQ-LLM is used to annotate a larger dataset. Finally, an open-source model is fine-tuned using the HQ-LLM-annotated data, optimizing both effectiveness and cost-efficiency.

Parameter	Value
LoRA Target	All
Per Device Train Batch Size	16
Gradient Accumulation Steps	4
LR Scheduler Type	Cosine
Warmup Ratio	0.1
Learning Rate	3e-05
Number of Train Epochs	5
Max Gradient Norm	1.0
Quantization Bit	4
LoRA+ LR Ratio	16.0
FP16	True
Validation Size	0.1

Table 8: Hyperparameters used for LoRA fine-tuning of models.

Category	Krippendorff's	Fleiss'
Threats to Culture or Identity	0.72	0.72
Threats to Survival or Physical Security	0.62	0.62
Vilification/Villainization	0.70	0.73
Explicit Dehumanization	0.68	0.65
None	0.70	0.73

Table 9: Inter-Annotator Agreement: Krippendorff's Alpha and Fleiss' Kappa for Russian war bloggers data.

Category	Cohen's	Accuracy	F1
Threats to Culture or Identity	0.83	0.92	0.92
Threats to Survival/Security	0.75	0.82	0.80
Vilification/Villainization	0.80	0.90	0.90
Explicit Dehumanization	0.85	0.94	0.94
None	0.80	0.92	0.92

Table 10: Inter-Annotator Agreement and Model Performance: Cohen's Kappa (Agreement), Accuracy, and F1 Score between majority vote and HQ-LLM (GPT-4o) on Russian war bloggers data. For F1 and Accuracy, five trials were ran, and with a temperature set to 0, identical results were obtained in every trial.

Category	Instance Counts
Threats to Culture or Identity	45
Threats to Survival or Physical Security	41
Vilification/Villainization	52
Explicit Dehumanization	32
None	87
Total Data Points	**212**

Table 11: Human-annotated gold set summary for Ukrainan war bloggers data.

Category	Krippendorff's	Fleiss'
Threats to Culture or Identity	0.75	0.78
Threats to Survival or Physical Security	0.77	0.76
Vilification/Villainization	0.78	0.79
Explicit Dehumanization	0.80	0.801
None	0.77	0.78

Table 12: Inter-Annotator Agreement: Cohen's Kappa for Ukrainian war bloggers data.

Category	Cohen's	Accuracy	F1 Score
Threats to Culture or Identity	0.80	0.90	0.91
Threats to Survival or Physical Security	0.76	0.81	0.82
Vilification/Villainization	0.78	0.89	0.88
Explicit Dehumanization	0.81	0.96	0.96
None	0.83	0.93	0.92

Table 13: Inter-Annotator Agreement and Model Performance: Cohen's Kappa (Agreement), Accuracy, and F1 Score between majority vote and HQ-LLM (GPT-4o) on Ukrainian war bloggers data. For F1 and Accuracy, five trials were ran, and with a temperature set to 0, identical results were obtained in every trial.

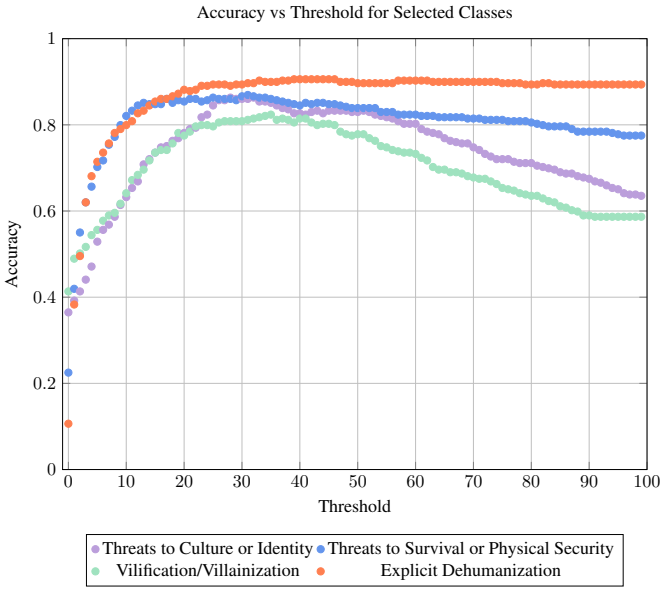


Figure 11: Scatter plot of accuracy vs. threshold for selected classes. This plot illustrates how adjusting confidence thresholds for classification logits affects accuracy across different classes in a new domain (Gab corpus), using a model initially fine-tuned on Russian war bloggers.

Table 14: Class Distribution and Split Summary (Percentages) - Russian War-Bloggers Dataset (7:2:1 split, 9,412 examples total)

Class	Overall (%)	Train (%)	Validation (%)	Test (%)
Threats to Culture or Identity	21.39	21.39	21.51	21.32
Threats to Survival or Physical Security	16.92	16.93	16.90	16.89
Vilification/Villainization	20.54	20.51	20.57	20.50
Explicit Dehumanization	3.93	3.92	3.90	3.90
None	54.98	55.00	54.96	55.05

Table 15: Class Distribution and Split Summary (Percentages) - Ukrainian War-Bloggers Dataset (7:2:1 split, 9,810 examples total)

Class	Overall (%)	Train (%)	Validation (%)	Test (%)
Threats to Culture or Identity	13.05	13.00	14.21	17.03
Threats to Survival or Physical Security	26.83	26.70	26.51	21.92
Vilification/Villainization	20.93	20.76	20.98	20.53
Explicit Dehumanization	5.71	5.73	5.34	4.76
None	47.82	48.62	48.43	51.91

Table 16: Class Distribution and Split Summary (Percentages) - Russian and Ukrainian War-Bloggers Dataset (7:2:1 split, 17,300 examples total)

Class	Overall (%)	Train (%)	Validation (%)	Test (%)
Threats to Culture or Identity	17.05	17.04	17.11	17.03
Threats to Survival or Physical Security	22.00	21.88	21.91	21.92
Vilification/Villainization	20.60	20.50	20.48	20.53
Explicit Dehumanization	4.76	4.76	4.67	4.76
None	51.48	51.92	51.91	51.91

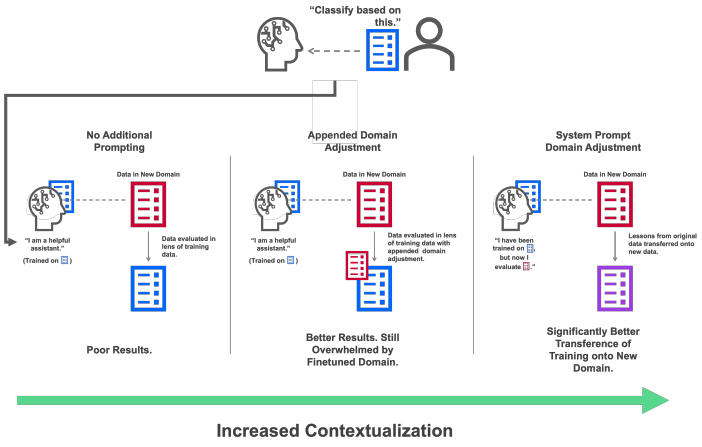


Figure 12: System Prompt Steering: demonstrates increased contextualization with the use of system prompt steering.

Table 17: Model Performance (Mean ± Std. Dev.): Accuracy and F1 Score for LLaMA3-8b-Instruct, Mistral, LLaMA2, and XLM-RoBERTa on Russian War Bloggers Data (7:2:1 split).

Model	**Category**	**Accuracy (%)**	**F1 Score (%)**
LLaMA3-8b-Instruct	Threats to Culture or Identity	0.92 ± 0.00	0.92 ± 0.00
	Threats to Survival or Physical Security	0.82 ± 0.00	0.80 ± 0.00
	Vilification/Villainization	0.90 ± 0.00	0.90 ± 0.00
	Explicit Dehumanization	0.94 ± 0.00	0.94 ± 0.00
	None	0.92 ± 0.00	0.92 ± 0.00
Mistral	Threats to Culture or Identity	0.89 ± 0.00	0.89 ± 0.00
	Threats to Survival or Physical Security	0.81 ± 0.00	0.80 ± 0.00
	Vilification/Villainization	0.89 ± 0.00	0.88 ± 0.00
	Explicit Dehumanization	0.91 ± 0.00	0.90 ± 0.00
	None	0.86 ± 0.00	0.85 ± 0.00
LLaMA2	Threats to Culture or Identity	0.89 ± 0.00	0.87 ± 0.00
	Threats to Survival or Physical Security	0.80 ± 0.00	0.80 ± 0.00
	Vilification/Villainization	0.88 ± 0.00	0.87 ± 0.00
	Explicit Dehumanization	0.90 ± 0.00	0.89 ± 0.00
	None	0.85 ± 0.00	0.85 ± 0.00
XLM-RoBERTa	Threats to Culture or Identity	0.59 ± 0.01	0.58 ± 0.01
	Threats to Survival or Physical Security	0.61 ± 0.01	0.60 ± 0.01
	Vilification/Villainization	0.64 ± 0.01	0.63 ± 0.01
	Explicit Dehumanization	0.67 ± 0.01	0.67 ± 0.01
	None	0.70 ± 0.02	0.69 ± 0.01

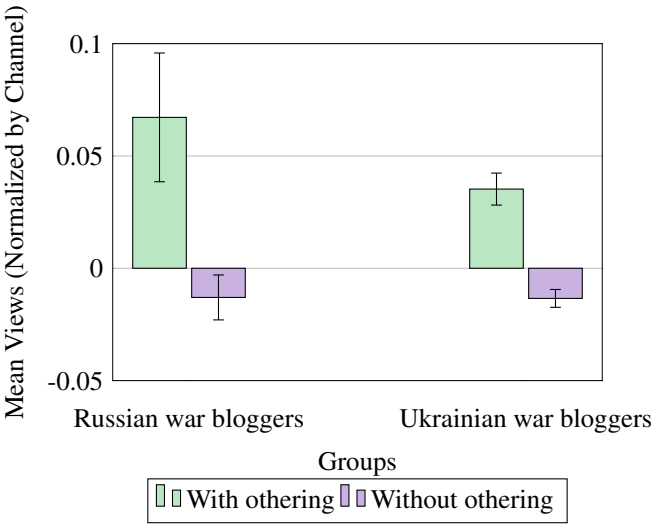


Figure 13: Comparison of mean views with and without othering (z-score channel-normalized) following crises. The bars represent the mean views for Russian and Ukrainian war bloggers, with and without othering, and the error bars indicate the standard error.

Table 18: Model Performance (Mean ± Std. Dev.): Accuracy and F1 Score for LLaMA3-8b-Instruct, Mistral, LLaMA2, and XLM-RoBERTa on Ukrainian War Bloggers Data (7:2:1 split).

Model	**Category**	**Accuracy (%)**	**F1 Score (%)**
LLaMA3-8b-Instruct	Threats to Culture or Identity	0.90 ± 0.00	0.91 ± 0.00
	Threats to Survival or Physical Security	0.81 ± 0.00	0.82 ± 0.00
	Vilification/Villainization	0.89 ± 0.00	0.88 ± 0.00
	Explicit Dehumanization	0.96 ± 0.00	0.96 ± 0.00
	None	0.93 ± 0.00	0.92 ± 0.00
Mistral	Threats to Culture or Identity	0.89 ± 0.00	0.88 ± 0.00
	Threats to Survival or Physical Security	0.79 ± 0.00	0.79 ± 0.00
	Vilification/Villainization	0.84 ± 0.00	0.84 ± 0.00
	Explicit Dehumanization	0.93 ± 0.00	0.92 ± 0.00
	None	0.88 ± 0.00	0.87 ± 0.00
LLaMA2	Threats to Culture or Identity	0.87 ± 0.00	0.86 ± 0.00
	Threats to Survival or Physical Security	0.81 ± 0.00	0.81 ± 0.00
	Vilification/Villainization	0.82 ± 0.00	0.83 ± 0.00
	Explicit Dehumanization	0.91 ± 0.00	0.90 ± 0.00
	None	0.86 ± 0.00	0.85 ± 0.00
XLM-RoBERTa	Threats to Culture or Identity	0.58 ± 0.01	0.57 ± 0.01
	Threats to Survival or Physical Security	0.62 ± 0.01	0.61 ± 0.01
	Vilification/Villainization	0.64 ± 0.01	0.66 ± 0.01
	Explicit Dehumanization	0.69 ± 0.01	0.68 ± 0.01
	None	0.71 ± 0.02	0.70 ± 0.01

Category	Instance Counts
Threats to Culture or Identity	120
Threats to Survival or Physical Security	74
Vilification/Villainization	136
Explicit Dehumanization	35
None	114
Total Data Points	**329**

Table 19: Summary of human annotations for Gab data.

	Dataset F1 Score		
Prompting Type	Russian	Ukrainian	Gab
No Additional Prompt	0.74	0.66	0.53
In-Context Learning	0.72	0.63	0.63
System Prompt	**0.78**	0.75	0.76
RDA	**0.78**	**0.76**	**0.77**

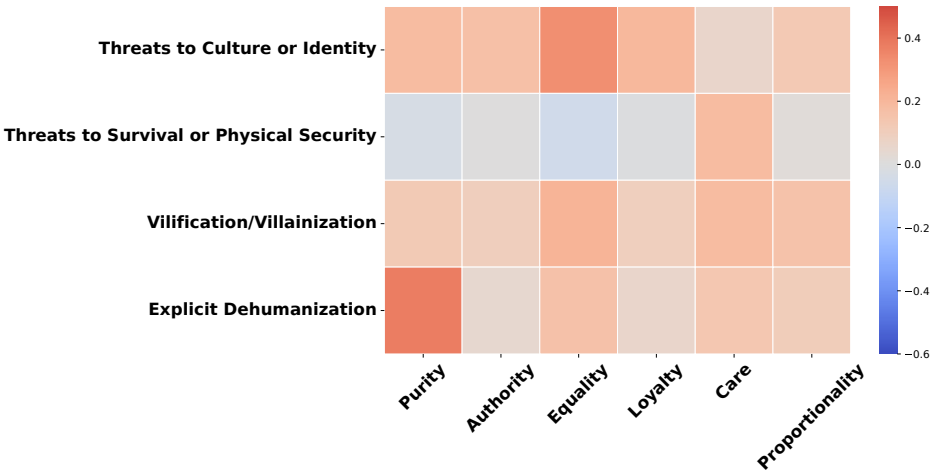
Table 21: F1-Score comparison across different prompting types and test sets (Russian Dataset, Ukrainian Dataset, Gab Dataset). Each test was run five times, and with a temperature set to 0, identical results were obtained in every trial.

Category	Cohen's
Threats to Culture or Identity	0.87
Threats to Survival or Physical Security	0.88
Vilification/Villainization	0.88
Explicit Dehumanization	0.92
None	0.91

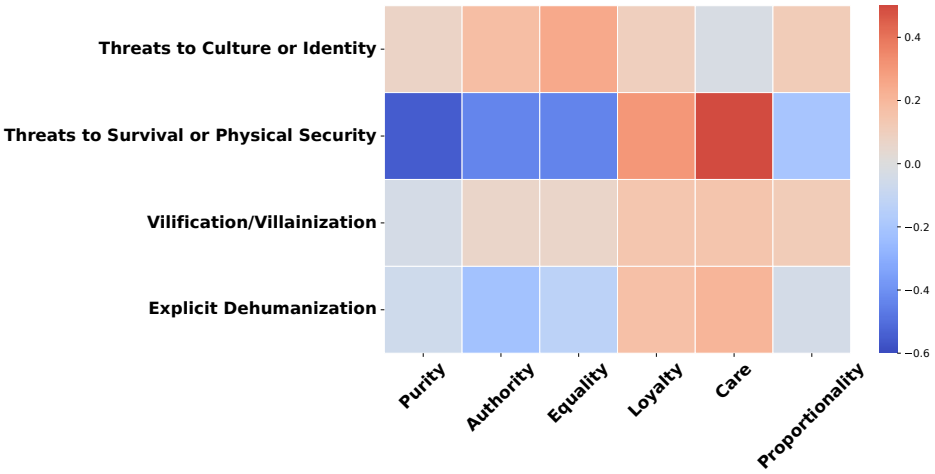
Table 20: Inter-Annotator Agreement: Cohen's Kappa for Gab data.

	Centrality Metric	
Community	Degree	Eigenvector
Russian	0.290 (+13.2%)	0.385 (+14.5%)
Ukrainian	0.177 (+32.1%)	0.136 (-7.8%)

Table 22: Centrality and othering messages following key events. Spearman correlation between a channel's proportion of messages with othering language and its degree and eigenvector centralities. All correlations are significant at the $p < 0.01$ level.



(a) Log-odds ratios for morality devices in **Russian war bloggers**' messages.



(b) Log-odds ratios for morality devices in **Ukrainian war bloggers**' messages.

Figure 14: Comparison of log-odds ratios for morality devices across othering categories in Russian and Ukrainian war bloggers' messages. The color intensity reflects the strength of the association, with warmer colors (red) indicating higher positive log-odds ratios and cooler colors (blue) representing negative or lower values.

Date	Event	Key
2022-02-08	Putin claims allowing Ukraine to join NATO would increase the prospects of a Russia-NATO conflict that could turn nuclear.	1a
2022-02-21	Putin cites Nazism in Ukraine in speech legitimizing upcoming invasion.	2a
2022-02-24	Russia invades Ukraine.	-
2022-04-19	Russia officially pivots to 'next phase' of war. Russia shifted its troops from the Kyiv offensive to Ukraine's eastern Donbas region, and the amassed forces launched a broad attack there on April 18. Ukraine called it a "new phase of the war."	3a
2022-06-01	The Biden administration authorizes an 11th presidential drawdown of security assistance to Ukraine valued at up to $700 million.	4a
2022-06-23	The Biden administration authorizes a 13th presidential drawdown of security assistance to Ukraine valued at up to $450 million.	5a
2022-07-08	The Biden administration announces $400 million in additional security assistance for Ukraine.	6a
2022-08-01	The Biden administration announces $550 million in additional security assistance for Ukraine.	7a
2022-09-28	United States Department of Defense announces approximately $1.1 billion in additional security assistance for Ukraine.	8a
2023-02-03	United States Department of Defense announces a significant new package of security assistance for Ukraine, including the authorization of a presidential drawdown of security assistance valued at up to $425 million, as well as $1.75 billion in Ukraine Security Assistance Initiative (USAI) funds.	9a

Table 23: Key events in the war discussed by Russian war bloggers. The "Key" column corresponds to the labeled vertical lines in Figure 4. Entries without a key were included in the data analysis but are not visualized due to their proximity to other points.

Date	Event	Key
2022-02-08	Putin claims allowing Ukraine to join NATO would increase the prospects of a Russia-NATO conflict that could turn nuclear.	1b
2022-02-21	Putin cites Nazism in Ukraine in speech legitimizing upcoming invasion.	2b
2022-02-24	Russia invades Ukraine.	-
2022-03-02	Russia captures Kherson.	-
2022-03-21	Russian troops used stun grenades and gunfire to disperse a rally of pro-Ukrainian protesters in the occupied southern city of Kherson on Monday.	3b
2022-03-21	Russia abandons Kherson.	-
2022-04-01	Reports of Russian atrocities in Bucha begin to surface.	-
2022-07-03	Russia captures Lysychansk, all of Luhansk Oblast	4b
2022-08-29	Ukraine launches first major counteroffensive.	5b
2022-09-21	Ukraine forces Russian retreat.	6b
2022-11-11	Ukraine recaptures Kherson.	7b
2022-12-29	Major Russian missile attack on infrastructure facilities in Kyiv, Kharkiv, Lviv, and other cities.	8b
2023-02-09	Russia launches second spring offensive.	9b

Table 24: Key events in the war discussed by Ukrainian war Bloggers. The "Key" column corresponds to the labeled vertical lines in Figure 4. Entries without a key were included in the data analysis but are not visualized due to their proximity to other points.

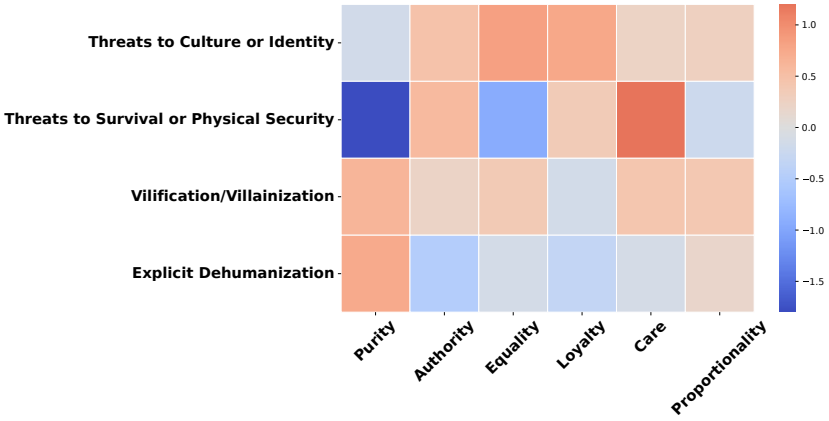


Figure 15: Heatmap displays the log-odds ratios for the use of various morality devices (Purity, Authority, Equality, Loyalty, Care, and Proportionality) across different othering categories (Threats to Culture or Identity, Threats to Survival or Physical Security, Vilification/Villainization, and Explicit Dehumanization) in the **Gab users'** messages. The color intensity reflects the strength of the association, with warmer colors (red) indicating higher positive log-odds ratios and cooler colors (blue) representing negative or lower values.