

A Multilingual Dataset for Dehumanisation Detection in Russian and Ukrainian Wartime Communication

Elizaveta Chernenko

Oxford Internet Institute
University of Oxford
United Kingdom

elizaveta.chernenko@oii.ox.ac.uk

Balazs Vedres

Department of Network and Data Science
Central European University
Austria

vedresb@ceu.edu

Abstract

Dehumanising language represents one of the most dangerous types of hate speech, that is known to hinder empathy, undermine prosocial behaviour, and contribute to between-group aggression. In wartime contexts, dehumanisation can escalate conflict by fostering support for war and enabling violence, extreme outgroup-targeted policies, and genocide. Yet, there is very limited data available to study online dehumanisation, and there are no respective datasets available in Russian and Ukrainian languages. We address this gap by introducing a novel dataset of dehumanising language from Telegram communication during Russia’s full-scale invasion of Ukraine. Using LLM-assisted annotation, we construct a corpus of 238,190 posts from leading Russian and Ukrainian news and political Telegram channels in 2022, classified as non-dehumanising, marginally dehumanising, or dehumanising. We additionally release a manually annotated evaluation set of 1,000 posts, labelled for six manifestations of dehumanisation, the targeted outgroup, and overall dehumanisation. Together, these resources provide a theory-informed foundation for computational research on dehumanisation in Russian- and Ukrainian-language wartime communication.

CONTENT WARNING: This paper contains some examples of abusive and hateful content that some readers may find offensive or distressing.

1 Introduction

The study focuses on *dehumanising language*, considered one of the most harmful forms of hostile communication which has been shown to hinder empathy (Bastian et al., 2011), undermine prosocial behaviour (Wohl et al., 2012), and contribute to between-group aggression (Viki et al., 2013). In wartime contexts, dehumanisation can

escalate conflict by fostering support for war (Jackson and Gaertner, 2010) and by enabling violence, extreme outgroup-targeted policies (Leidner et al., 2013; Maoz and McCauley, 2008), and genocide (Haslam, 2006; Kuchenbuch, 2016).

In this paper, we introduce a theory-informed dataset of dehumanising language in wartime online communication, collected from popular Russian and Ukrainian Telegram channels. We contribute a large LLM-annotated corpus alongside a manually annotated evaluation set capturing both overall dehumanisation and its specific manifestations.

Dehumanisation is understood as “the act of perceiving or treating people as less than human” (Haslam and Stratemeyer, 2016). It has been studied across a wide range of contexts, including the dehumanisation of racial and ethnic minorities, political groups and migrants, gender-based and other perceived outgroups, such as people with disabilities, criminals, perpetrators and victims (Haslam, 2006; Schwartz and Struch, 1989).

It is important to distinguish dehumanising language from other types of hate speech. The latter generally refers to various forms of harmful or hostile communication targeted at individuals or groups, typically expressed through insults, threats, or negative attitudes. In contrast, dehumanising language represents a particular form of intolerant and hostile language that operates through framing of an outgroup as less than human, thereby constituting an extreme form of othering and moral exclusion (Engelmann et al., 2024; Gerard et al., 2025). Importantly, dehumanising language represents a threat to democracy (Dodeigne, 2025).

While this body of research is growing, dehumanising language has not been analysed enough as other types of hate speech, such as misogyny, racism, homotransphobic language, etc. (Basile et al., 2019; Chakravarthi, 2024; Hartvigsen et al., 2022; Dorn et al., 2024)

Importantly, dehumanisation is not a single, uniform phenomenon, but a multifaceted and multi-component psychological construct that can manifest in different forms and degrees, and is typically detected via “proxies” (Caporusso et al., 2024; Landry and Seli, 2024). It may take the form of explicit comparison to non-human entities (e.g., animals or objects), but can also emerge through a combination of more subtle dimensions, including deindividuation (Haslam, 2006), denial of agency or emotional capacity, denial of prosocial values (Kelman, 1973), extreme negative evaluation and stereotypes (Bruneau et al., 2022; Haslam, 2006; Lees and Cikara, 2019; Kteily et al., 2015; Tsfaty and Cohen, 2005), and expressions of moral disgust or indifference (Caporusso et al., 2024; Saffari et al., 2025).

Dehumanisation is typically accompanied by affective polarisation, whereby strong positive identification with the ingroup is accompanied by increasingly negative perceptions of the outgroup (Iyengar et al., 2012), as well as by perceptions of outgroup members as fundamentally different from the ingroup, closely linking the phenomenon to psychological distancing and othering (Mendelsohn et al., 2020; Gerard et al., 2025).

Thus, dehumanisation is not reducible to surface-level toxicity: it reflects an underlying process of denying humanness, which may or may not be accompanied by overtly offensive language (Engelmann et al., 2024). As a result, standard hate speech or toxicity classifiers, typically tuned to detect explicit insults or profanity, are not well-suited to reliably capture dehumanising content, particularly in its implicit or metaphorical forms (Saffari et al., 2025; Assenmacher et al., 2025).

The proposed dataset is specifically designed to capture instances of dehumanisation directed at nationally or politically defined outgroups, given their particularly severe consequences, including facilitating intergroup violence and legitimising harm.

2 Related work

Dehumanisation has long been studied in psychology and communication research as a multidimensional process through which individuals or social groups are denied humanness (Haslam, 2006; Haslam et al., 2008; Kteily and Landry, 2022). Early approaches primarily focused on overt manifestations of dehumanisation, particularly explicit

non-human comparisons and animal metaphors.

Computational work on dehumanisation, however, has emerged only recently. Mendelsohn et al. (2020) presented one of the first computational frameworks for measuring dehumanisation, using lexicon-based features, word embeddings and connotation frames to operationalise four theoretically motivated dimensions of dehumanisation: negative evaluation, denial of agency, moral disgust, and vermin metaphors, in a large corpus of New York Times articles. While influential, this approach measures dehumanising attitudes indirectly and is primarily intended for aggregate analyses rather than identifying dehumanising expressions in individual texts.

More recent work has shifted towards developing annotated datasets and supervised models for automatic detection of dehumanisation expressions. Vidgen et al. (2021) introduced a dynamically generated English hate speech dataset comprising over 41,000 instances annotated for multiple hate speech categories, including 906 examples labelled as dehumanisation. However, the dataset contains synthetic rather than naturally occurring content. Engelmann et al. (2024) contributed one of the first datasets dedicated specifically to dehumanisation, consisting of a manually annotated evaluation set of 918 English texts accompanied by a much larger automatically collected corpus of approximately 565,000 candidate texts extracted from OpenSubtitles and Common Crawl. However, the annotation task is limited to binary dehumanisation classification and does not distinguish different manifestations of dehumanisation or provide target-group annotations. Joshi (2024) proposed an English Reddit dataset designed to capture the contextual interpretation of potentially dehumanising expressions. Using more than 10,000 Reddit discussion threads and a curated lexicon of animalistic metaphors, the study demonstrates that contextual information is essential for distinguishing genuinely dehumanising uses from benign mentions of the same lexical items. Further, Assenmacher et al. (2025) introduced a theory-informed bilingual English-German dataset collected from Twitter and Reddit, comprising approximately 16,000 crowdworker-annotated instances covering multiple dimensions of dehumanisation, including more subtle forms beyond explicit animalistic metaphors. They further released a large automatically collected pool of over 490,000 candidate dehumanising texts and showed that mod-

els fine-tuned on their corpus outperform existing approaches in both zero-shot and few-shot settings. [Mendelsohn and Budak \(2025\)](#) similarly extended computational work on dehumanisation by analysing metaphor-based dehumanisation of migrants in approximately 400,000 U.S. tweets on immigration and releasing a manually annotated evaluation dataset of 1,600 posts.

Parallel work has examined hate speech in the context of the Russia–Ukraine war, but has focused on English-language content and broader forms of hate speech rather than dehumanisation specifically. [Bhandari et al. \(2023\)](#) introduced a benchmark dataset for detecting hate speech and identifying its targets in English-language text embedded in images related to the Russia–Ukraine war. This dataset was later used in the CASE 2024 shared task ([Thapa et al., 2024b](#)). Building on this work, [Thapa et al. \(2024a\)](#) presented the RUHate-MM dataset, a manually annotated corpus of 20,675 English-language Twitter/X posts combining textual and visual information and annotated both for hate speech and for the type of target.

Several recent studies have also examined dehumanising discourse and related phenomena in Russian-language social media during the Russia–Ukraine war. [Burovova and Romanyshyn \(2024\)](#) analysed dehumanising dynamics in Russian online discourse using computational indicators derived from [Mendelsohn et al. \(2020\)](#). However, their approach relies on indirect lexical proxies for individual dimensions of dehumanisation – such as denial of agency or extreme negative evaluation – rather than on validated ground-truth annotations of dehumanising language. [Gerard et al. \(2025\)](#) investigate the closely related phenomenon of outgroup othering in Russian and Ukrainian Telegram war-blogger channels using large language models. They demonstrate that existing hate speech and fear speech classifiers fail to capture many forms of subtle, context-dependent othering, and show that although othering overlaps with hate speech, the two phenomena are not reducible to one another. Importantly, their analysis indicates that dehumanising language tends to be more toxic on average and less frequently accompanied by moral framing than othering, highlighting the need to distinguish these concepts computationally.

Taken together, the existing literature reveals three important gaps. First, nearly all publicly available dehumanisation datasets are limited to English, with only one recent resource extending

to German, and draw on platforms such as Twitter/X, Reddit, Common Crawl, and OpenSubtitles. Second, computational resources related to the Russia–Ukraine war primarily target other forms of hate speech. Third, the only existing study of Russian-language wartime dehumanisation relies on indirect measures of dehumanising attitudes rather than expert-validated ground-truth annotations.

Contributions We address these gaps by introducing the first theory-informed, expert-validated resource for studying dehumanising language in Russian and Ukrainian. It comprises a manually annotated evaluation set of 1,000 Telegram posts, labelled for six manifestations of dehumanisation, targeted outgroups, and overall dehumanisation (classified as non-dehumanising, marginally dehumanising, or dehumanising), alongside a validated LLM-annotated corpus of 238,190 posts classified into the same three categories. By focusing on Telegram, the principal communication platform in the two countries at war, and grounding annotation in psychological theories of dehumanisation, the dataset provides a resource for computational research on dehumanisation detection as well as broader studies of wartime communication, propaganda, and intergroup conflict.

3 Dataset

Telegram channels are among the most widely used platforms for accessing and discussing war-related news in both Russia ([Schechner, 2022](#)) and Ukraine ([Chernenko and Dutton, 2025](#)), particularly after the 2022 full-scale invasion. In Ukraine, about one-fifth of users joined the platform after the invasion ([Zarembko et al., 2024](#)), and nearly half now cite it as a primary news source ([Polikovska, 2024](#)). At the time of writing, there are approximately one hundred thousand Telegram channels in Ukraine and over one million in Russia ([TGStat, 2025](#)). Their appeal stems from a combination of low moderation, intuitive and convenient interface, and real-time broadcasting ([Hoskins and Shchelin, 2023](#)), which positions Telegram as a central arena for public opinion formation and information warfare in the region ([Chernenko and Dutton, 2025](#); [Hanley and Durumeric, 2024](#)).

This outstanding role of Telegram channels, and particularly its unmoderated nature, motivated our choice of this platform as a source for studying wartime outgroup dehumanisation.

3.1 Data Collection and Pre-Processing

We selected the 89 most popular Russian and 95 most popular Ukrainian news and political Telegram channels, identified using TGStat – the largest catalogue of Telegram channels and groups in the region (TGStat, 2025). Specifically, we have chosen two categories of Telegram channels as they are classified by the catalogue: general “News and Media” and “Political” channels and collected all of the messages published between December 24, 2021, and April 24, 2022. This period includes two months before and two months after the full-scale Russian invasion of Ukraine on February 24, 2022, allowing the capture of both anticipatory and reactive discourse, as well as potential information operations across the conflict timeline.

The original dataset included 1,025,797 posts. We first cleaned it from posts with no text (192,374 multimedia-only posts). Further, posts that did not mention both national groups of interest (595,219 cases) – Ukrainians and Russians – were excluded from further analysis. This was done to create a dataset for studying intergroup dynamics. We applied a filtering process using a dictionary of group-specific identifiers¹, including formal (e.g., ‘Kremlin’), positional (e.g., ‘enemy’), and derogatory or dehumanising terms (e.g., ‘Ukrofascists’, ‘orc’). Hashtags and channel titles were excluded from matching, ensuring that posts mentioning group identifiers only in hashtags or channel names were retained only if relevant identifiers also appeared in the main post text. The full set of dictionaries with 922 tokens applied for filtering is available in Chernenko (2026)

Finally, we removed all posts in languages other than Ukrainian or Russian (14 cases). The resulting dataset includes 238,190 posts.

3.2 Concept Operationalisation

Following Haslam and Stratemeyer (2016), we define dehumanisation as treating the other as less than human. In its most explicit forms, dehumanisation manifests itself in non-human comparisons. However, it may also be present through other manifestations. Therefore, in line with literature discussed in Section 1, outgroup dehumanisation was

operationalised as the presence of at least one of the following six manifestations²:

1. Explicit comparison to an animal, machine, non-human creature or inanimate substance
2. Deindividuated, entitative outgroup perception
3. (Moral) disgust or indifference
4. Extreme negative evaluation
5. Denial of agency: portraying outgroup members as incapable of emotion, rationality, or independent decision-making
6. Denial of prosocial values and communality – depicting the outgroup as incapable of mutual care or social belonging.

Table 1 summarises this conceptual framework. To illustrate dehumanising metaphors and discursive framing, we provide several examples below. Words frequently used in posts dehumanising Ukrainians include terms that associate Ukrainians or the Ukrainian army with Nazis or fascists (e.g., ‘Ukrofascisty,’ ‘Ukronazisty,’ ‘naziki’). Other expressions are derogatory neologisms formed by altering words related to Ukrainian national identity (e.g., ‘ukrop,’ literally ‘dill’; ‘ukry,’ an abbreviated and pejorative form of ‘Ukrainians’) or national symbols (e.g., ‘vyshivata,’ derived from ‘vyshivanka,’ a traditional embroidered shirt, and incorporating a non-human metaphor likening Ukrainians to cotton, ‘vata’). Such words and metaphors have also been documented in prior research (Alyukov et al., 2023).

Posts dehumanising Russians likewise frequently employ terms that associate Russians or the Russian army with fascism or Nazism (e.g., ‘russhisty,’ ‘russhizm,’ ‘Z-nazisty’) and derogatory forms derived from references to Russian national identity (e.g., ‘rusnya,’ that refers to Russians, but sounds like some entitative, homogenous and unpleasant substance). In addition, they include expressions that compare the Russian outgroup to animals such as pigs or cattle (e.g., ‘svonosobaki’ – ‘pig-dogs’; ‘svinorasha’ – ‘pig-Russia’; ‘skotorusy’ – ‘cattle-Russians’) or to mythical or folkloric creatures associated with evil, aggression, and disgust (e.g., ‘orcs,’ ‘Mordor’).

The full codebook for manual annotation is provided in Appendix A. It is supported by dictionary

¹All of these dictionaries, except for positional keywords dictionary for Russian Telegram result in Precision and Recall of at least 95% (positional keywords dictionary for Russian results in approximately 90% precision). Evaluation was done based on randomly chosen 100 posts classified as ‘positive’ and 100 classified as ‘negative’.

²Note that we do not utilise othering as a dimension of dehumanisation. While these phenomena are related, findings show (Gerard et al., 2025), that dehumanising language tends to be more toxic and less related to moral framing than othering, which points to the need to distinguish the two concepts computationally.

Table 1: Types of dehumanising manifestations, indicative language, and examples

DH manifestations	Indicative keywords	Communicated Meaning	Sample Examples
Explicit comparison to non-human entities	Words like “orc”, “vata”; death/killing euphemisms referring to people	“He is an animal”, “They’re not people”, “They’re consumables”	“Зеленский признал — тысячи ВСУшников списаны (Zelensky admitted that thousands of AFU soldiers have been written off)”; “Американцы навязывают своим питомцам террористическую тактику (Americans are foisting terrorist tactics on their pets)”; “Гниющие нацисты (Rotting Nazis)”
Deindividualised perception	Words like “ukry”, “rusnya”; explicit claims of sameness	“They are all the same”	“Это у рашистов нет жертв? Им просто наплевать на свою солдатню (So, the Russists have no casualties? They simply don’t care for their soldiery)”
(Moral) disgust or indifference	Words like “khokhly”, “moscal”, “nazik”	“They are pathetic creatures”, “They are disgusting”	“Блеск и нищета украинской пропаганды (The Splendour and Misery of Ukrainian Propaganda)”
Extreme negative evaluation	Words like “orda”, “banderovets”; strong insults	“They are terrible”	“Живодёр из Азова (Azov knacker)”; “Борзые укронацисты (Brazen Ukronazis)”
Denial of agency	Words like “maidaun”, “malorossiya”; explicit expressions of incapacity	“He is incapable of reason”, “She has no free will”	“Марионеточный режим... множатся на ноль (Puppet regime... They’re being reduced to zero)”
Denial of prosocial values	Words like “nazist”, “russhist”; explicit pointing to lack of morality	“Not capable of empathy”, “They have no meaningful social bonds to anyone”	“Рашисты кинули своих умирать (The Russists abandoned their own to die)”; “Расправы украинских солдат над российскими пленными (Ukrainian soldiers’ atrocities against Russian POWs)”

ies of 108 indicative dehumanising and derogatory keywords frequently used in posts dehumanising the two national groups (Chernenko, 2026).

3.3 Data Annotation

Building on the codebook, pilot annotation, iterative discussions between the two expert coders, we developed an initial labelled dataset of 222 Telegram posts, manually classified into three categories: non-dehumanising, marginally dehumanising, and dehumanising. Importantly, this initial annotated sample was *not intended to be statistically representative* of the full dataset, but served an instrumental purpose to develop, refine, and evaluate the classification approach (i.e., the combination of prompt and model). To ensure that the classifier captures the diversity of dehumanising expressions, dehumanising cases were deliberately over-represented in the sample.

The sample was constructed to ensure sufficient coverage across categories, with at least 20 cases for each dehumanisation category, language, and country (see Table 2). The sample includes posts from a wide range of channels across both countries, in both languages, and reflects variation in tone, format, and communicative context. As detailed in the annotation guidelines (Appendix A), posts containing explicit comparisons to non-human entities were automatically classified as dehumanising. In the absence of such comparisons, classification was based on the presence of other

Table 2: Composition and Diversity of the Originally Labelled Sample

Category	Share (%)	Channels (n)
Non-dehumanising cases	28	39
Marginally dehumanising cases	21	15
Dehumanising cases	51	42

dimensions of dehumanisation. Posts exhibiting none of the five additional dimensions were coded as non-dehumanising. For the remaining cases, classification relied on the overall evaluative judgment of a coder, informed by the combination and intensity of observed features. Posts that conveyed a strong cumulative impression of the outgroup as less than human, typically reflecting multiple (three or more) dimensions, were labelled as dehumanising. Posts exhibiting weaker signals (typically one or two dimensions) were categorised as marginally dehumanising. Annotation was conducted independently by two coders, achieving an inter-annotator agreement of 95 percent.

3.4 Model Choice and Prompt Engineering

To identify the most effective approach for classifying dehumanising language, we evaluated multiple prompt configurations and LLMs via the OpenRouter platform (OpenRouter Inc., 2023). Prompt experiments varied the temperature and compared zero-shot classification with configurations that added or removed: (1) a definition of dehuman-

Table 3: Post-classification performance evaluation

Setting	Accuracy	Macro F1	Macro Precision	Macro Recall
Three-level dehumanisation	0.780	0.722	0.761	0.725
Binary dehumanisation	0.836	0.830	0.826	0.855

isation; (2) contextual information, including the wartime Telegram setting, channel country, and post language; (3) a system-role description; (4) instructions and posts presented in different languages; (5) the annotation codebook and relevant dictionaries; (6) up to five labelled examples for each combination of dehumanisation category, country, and language; and (7) a request to explain the assigned label. Interestingly, adding labelled examples consistently reduced classification performance. This suggests that, given the heterogeneous manifestations of dehumanisation, a small set of examples may narrow the model’s interpretation towards the specific forms represented in those examples rather than improving generalisation across the construct. The final prompt was written in English while preserving each Telegram post in its original language. It included the definition of dehumanisation, a system-role description, contextual information, classification categories and decision rules, and a summary of the annotation codebook and dictionaries, and requested a brief explanation of the assigned label. We used a temperature of 0.075 to optimise classification consistency.

We subsequently evaluated multiple LLMs using this prompt, including:

- OpenAI: GPT-4o-mini 1.63B, 30.6B
- DeepSeek: DeepSeek V3 19.8B
- Meta: LLaMA 3.3 3.8B, LLaMA 3.3 70B Instruct
- Mistral AI: Mistral 7B
- Alibaba: Qwen 2 7B
- Microsoft: Phi 3 3.8B, 14B
- Google: Gemma 2 9B, 27B
- Cohere: Command R 35B

We selected Macro F1 as the primary metric for model selection because the class distribution is highly imbalanced, with non-dehumanising posts constituting the majority³. According to this cri-

³Under such conditions, accuracy can be misleading: a classifier favouring the majority class may achieve high accuracy while performing poorly on the substantively important dehumanising categories. Macro F1 gives equal weight to performance across classes and is therefore better aligned with the objectives of this study.

terion, Meta’s **LLaMA 3.3 70B Instruct** was the best-performing model. Point estimates and bootstrapped confidence intervals for all evaluated models, together with a dictionary-based baseline, are reported in Table 4. On the manually annotated dataset, LLaMA 3.3 70B Instruct achieved Accuracy = 0.810, Precision = 0.804, Recall = 0.810, and F1-score = 0.807.⁴ GPT-4o-mini showed the closest performance. Although the difference between the two models was not statistically conclusive, as their confidence intervals overlapped, LLaMA 3.3 70B Instruct was selected based on its slightly higher Macro F1 and greater financial feasibility for large-scale annotation.

Finally, the chosen classifier was applied to the large dataset of 238,190 posts. To further increase reliability, each post was annotated twice using independent LLM runs. Posts that were consistently classified as either *dehumanising* or *non-dehumanising* were retained in the dataset with that label. Posts that received inconsistent classifications or were labelled as marginally dehumanising in either run were included under the *marginally dehumanising* category. This procedure produced the final labelled dataset described below.

4 Results

The final dataset comprises 238,190 Telegram posts: 16,440 dehumanising, 89,879 marginally dehumanising, and 131,871 non-dehumanising. Of these, 176,217 are in Russian and 61,973 in Ukrainian.

Post-Classification Validation To evaluate classification performance, we validated the approach on a random sample of 1,000 posts from the full dataset: 719 in Russian and 281 in Ukrainian. Unlike the initial annotated sample, which was designed to maximise conceptual diversity rather than representativeness, this random validation sample provides an unbiased assessment of classification performance on the full dataset.

Each post was manually annotated for the six dimensions of dehumanisation (Section 3.2) and

⁴Precision, Recall, and F1 are macro-averaged.

Further analysis using this dataset, including the temporal dynamics of dehumanisation and the relationship between outgroup dehumanisation and ingroup cohesion, is presented in [Chernenko \(2025\)](#).

5 Conclusion

This study advances the understanding of outgroup dehumanisation in wartime online communication by introducing a carefully constructed and validated dataset of dehumanising and non-dehumanising Telegram posts from Russian and Ukrainian news and political channels during the early months of Russia's full-scale invasion of Ukraine. It contributes two complementary resources: a manually annotated ground-truth dataset of 1,000 posts and a larger automatically annotated corpus of 238,190 posts, each classified into three categories: non-dehumanising, marginally dehumanising, and dehumanising. To our knowledge, this is the first expert-validated computational resource for studying dehumanisation in Russian and Ukrainian languages, addressing an important gap in existing multilingual resources, which have focused predominantly on English language and on other forms of hate speech.

Beyond the dataset itself, the study contributes an operationalisation of dehumanisation grounded in theory, allowing the phenomenon to be distinguished from general hate speech, toxicity, and offensive language. By combining manual annotation, theory-informed codebook development, and large language model-assisted classification, we demonstrate a practical approach to identifying a complex, multifaceted, and context-dependent phenomenon at scale. Although the classifier achieves only moderate overall performance, validation suggests that it reliably distinguishes clear cases of dehumanisation from non-dehumanising communication, making it suitable for downstream analysis of wartime discourse.

The resulting resource provides a foundation for future computational and interdisciplinary research on dehumanisation, propaganda, intergroup conflict, and online political communication. It may also support the development and evaluation of more theoretically grounded dehumanisation detection systems, an area that remains substantially underdeveloped compared with general hate

tionalist leader whose faction initially collaborated with Nazi Germany in 1941 in pursuit of Ukrainian independence, but who was subsequently repressed and imprisoned by the Nazi regime after it rejected Ukrainian independence.

speech detection.

6 Ethical considerations

This dataset contains naturally occurring wartime communication, including messages expressing dehumanising attitudes toward nationally and politically defined outgroups. We recognise that releasing such data entails ethical risks, including the possibility that the content could be misused to reproduce and amplify harmful narratives. However, the dataset has been developed to support research on the detection, measurement, and mitigation of dehumanisation, and we believe that its scientific and societal benefits outweigh these risks.

The dataset is intended exclusively for academic research and is accompanied by documentation describing its theoretical foundations, annotation procedure, and appropriate use. As the data originate from publicly accessible Telegram channels, no private communications were collected or analysed. Finally, we acknowledge that the annotation of dehumanisation involves subjective judgement despite the use of a detailed codebook, multiple annotators, and consensus coding. We therefore encourage users to interpret the labels as theory-informed expert annotations rather than absolute ground truth.

7 Limitations and Future Work

This study has several limitations that suggest directions for future research.

First, the performance evaluation of the dehumanisation classifier is constrained by the size and design of the annotated samples. The initial model and prompt were chosen using a relatively small set of 222 posts, selected to maximise conceptual diversity rather than representativeness. The limited size of this initial annotated dataset remains an important limitation. Performance was subsequently evaluated on a random validation sample of 1,000 posts drawn from the full dataset of 238,190. Although larger validation samples would be desirable, the chosen size reflects a trade-off between statistical precision and the considerable complexity of expert annotation. Despite the relatively low prevalence of dehumanising content, the validation sample contained 98 dehumanising and 253 marginally dehumanising posts, providing a minimally sufficient basis for evaluating classifier performance.

Second, while the achieved performance is mod-

erate, it is comparable to – or exceeds – that reported for similarly complex classification tasks in relatively low-resource languages⁶. It is also consistent with the current state of the art in dehumanisation detection, where the phenomenon is treated as a multifaceted construct and even advanced models struggle with its implicit and context-dependent nature (Engelmann et al., 2024).

Third, reliance on LLM-based annotation introduces the risk of systematic bias (Baumann et al., 2025). To mitigate this, future applications of the dataset should consider implementing Design-based Supervised Learning (DSL) (Egami et al., 2023). By comparing model predictions with labels from the manually annotated validation sample, DSL enables systematic annotation errors (e.g., over- or under-classification) to be estimated and propagated into downstream statistical analyses, thereby improving the validity of inference.

A further extension would be to classify the six individual dimensions of dehumanisation underlying the three-category annotation scheme. This could provide a more fine-grained account of how different manifestations of dehumanisation occur and vary across contexts. However, reliable dimension-level classification would require substantial additional annotation and validation, particularly because individual dimensions may be more difficult to detect than the aggregate categories. Developing and validating such a multi-dimensional classifier therefore falls beyond the scope of the present dataset paper and represents a substantial direction for future research.

Finally, the classification of channel types relies on the TGStat taxonomy, which contains occasional inaccuracies. In particular, some channels associated with Russian influence operations targeting Ukraine are classified as Ukrainian, potentially obscuring the true ingroup–outgroup orientation used to identify outgroup dehumanisation. Future work should therefore refine channel-level metadata through manual verification or alternative classification approaches.

References

Maxim Alyukov, Maria Kunilovskaya, and Andrei Semenov. 2023. [Wartime media monitor \(WarMM-2022\): A study of information manipulation on Russian social media during the Russia-Ukraine war](#). In

Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 152–161, Dubrovnik, Croatia. Association for Computational Linguistics.

Dennis Assenmacher, Paloma Piot, Katarina Laken, David Jurgens, and Claudia Wagner. 2025. [Beyond the explicit: A bilingual dataset for dehumanization detection in social media](#). *Preprint*, arXiv:2510.18582.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 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*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Brock Bastian, Simon M. Laham, Sam Wilson, Nick Haslam, and Peter Koval. 2011. [Blaming, praising, and protecting our humanity: The implications of everyday dehumanization for judgments of moral status](#). *British Journal of Social Psychology*, 50(3):469–483.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del-Arco, Johannes B. Gruber, and Dirk Hovy. 2025. [Large language model hacking: Quantifying the hidden risks of using llms for text annotation](#). *Preprint*, arXiv:2509.08825.

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. [Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1994–2003.

Emile Bruneau, Andrés Casas, Boaz Hameiri, and Nour Kteily. 2022. [Exposure to a media intervention helps promote support for peace in Colombia](#). *Nature Human Behaviour*, 6(6):847–857.

Kateryna Burovova and Mariana Romanysyn. 2024. [Computational Analysis of Dehumanization of Ukrainians on Russian Social Media](#). In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 28–39, St. Julians, Malta. Association for Computational Linguistics.

Jaya Caporusso, Damar Hoogland, Mojca Brglez, Boshko Koloski, Matthew Purver, and Senja Polak. 2024. [A Computational Analysis of the Dehumanisation of Migrants from Syria and Ukraine in Slovene News Media](#). *arXiv preprint*.

⁶For example, Christner et al. (2025) report a macro-F1 score of 0.68 for disinformation detection.

- Bharathi Raja Chakravarthi. 2024. [Detection of homophobia and transphobia in YouTube comments](#). *International Journal of Data Science and Analytics*, 18(1):49–68.
- Elizaveta Chernenko. 2025. [Dehumanising Language as a Tool of Ingroup Cohesion – Case of Wartime Dehumanisation in Russian and Ukrainian Telegram](#). In *Proceedings of the 2nd International Workshop on Diffusion of Harmful Content on Online Web (DHOW '25)*, pages 44–51, Dublin, Ireland. ACM.
- Elizaveta Chernenko. 2026. [Dataset and resources for “a multilingual dataset for dehumanisation detection in russian and ukrainian wartime communication”](#).
- Elizaveta Chernenko and William H. Dutton. 2025. [Who Trusts Telegram in Wartime Ukraine? The Dynamics of Trust and the Use of Messenger-Based Social Media](#). *Ukrainian Analytical Digest*, 16:14–19.
- Clara Christner, Mykola Makhortykh, and Teresa Gil-Lopez. 2025. [Populist radical-right attitudes, media trust, and social media reliance: Combining survey and tracking data to investigate predictors of online exposure to disinformation](#). *Telematics and Informatics*, 98:102250.
- Jeremy Dodeigne. 2025. [Hostility and democratic erosion: Introducing the Political Hostility Scale \(PHOS\)](#). *Democratization*, pages 1–24.
- Rebecca Dorn, Lee Kezar, Fred Morstatter, and Kristina Lerman. 2024. [Harmful speech detection by language models exhibits gender-queer dialect bias](#). In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, Eaamo '24, New York, NY, USA. Association for Computing Machinery.
- Naoki Egami, Musashi Hinck, Brandon Stewart, and Hanying Wei. 2023. [Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 68589–68601. Curran Associates, Inc.
- Paul Engelmann, Peter Trolle, and Christian Hardmeier. 2024. [A dataset for the detection of dehumanizing language](#). In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 14–20, St. Julian's, Malta. Association for Computational Linguistics.
- Patrick Gerard, Tim Weninger, and Kristina Lerman. 2025. [Fear and Loathing on the Frontline: Decoding the Language of Othering by Russia-Ukraine War Bloggers](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 615–635.
- Hans W. A. Hanley and Zakir Durumeric. 2024. [Partial Mobilization: Tracking Multilingual Information Flows amongst Russian Media Outlets and Telegram](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 18:528–541.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. [ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.
- Nick Haslam. 2006. [Dehumanization: An Integrative Review](#). *Personality and Social Psychology Review*, 10(3):252–264.
- Nick Haslam, Stephen Loughnan, Yoshihisa Kashima, and Paul Bain. 2008. [Attributing and denying humanness to others](#). *European Review of Social Psychology*, 19(1):55–85.
- Nick Haslam and Michelle Stratemeyer. 2016. [Recent research on dehumanization](#). *Current Opinion in Psychology*, 11:25–29.
- Andrew Hoskins and Pavel Shchelin. 2023. [The war feed: Digital war in plain sight](#). *American Behavioral Scientist*, 67(3):449–463.
- Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. 2012. [Affect, Not Ideology: A Social Identity Perspective on Polarization](#). *Public Opinion Quarterly*, 76(3):405–431.
- Lydia Eckstein Jackson and Lowell Gaertner. 2010. [Mechanisms of moral disengagement and their differential use by right-wing authoritarianism and social dominance orientation in support of war](#). *Aggressive Behavior*, 36(4):238–250.
- Devavrat Joshi. 2024. [Decoding Dehumanization: Leveraging NLP to Identify Dehumanizing Language and Its Targets](#). In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, pages 87–96, Okayama Japan. ACM.
- Herbert G. Kelman. 1973. [Violence without Moral Restraint: Reflections on the Dehumanization of Victims and Victimizers](#). *Journal of Social Issues*, 29(4):25–61.
- Nour Kteily, Emile Bruneau, Adam Waytz, and Sarah Cotterill. 2015. [The ascent of man: Theoretical and empirical evidence for blatant dehumanization](#). *Journal of Personality and Social Psychology*, 109(5):901–931.
- Nour S. Kteily and Alexander P. Landry. 2022. [Dehumanization: Trends, insights, and challenges](#). *Trends in Cognitive Sciences*, 26(3):222–240.
- David Kuchenbuch. 2016. [In Search of the “Human Scale”: Delimiting the Social in German and Swedish Urban Planning in the 1930s and 1940s](#). *Journal of Urban History*, 42(6):1044–1064.

- Alexander P Landry and Paul Seli. 2024. A family-resemblances framework for dehumanization research. *Current Research in Ecological and Social Psychology*, 6:100185.
- Jeffrey Lees and Mina Cikara. 2019. Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*, 4(3):279–286.
- Bernhard Leidner, Emanuele Castano, and Jeremy Ginges. 2013. Dehumanization, Retributive and Restorative Justice, and Aggressive Versus Diplomatic Intergroup Conflict Resolution Strategies. *Personality and Social Psychology Bulletin*, 39(2):181–192.
- Ifat Maoz and Clark McCauley. 2008. Threat, Dehumanization, and Support for Retaliatory Aggressive Policies in Asymmetric Conflict. *Journal of Conflict Resolution*, 52(1):93–116.
- Julia Mendelsohn and Ceren Budak. 2025. When People are Floods: Analyzing Dehumanizing Metaphors in Immigration Discourse with Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8079–8103, Vienna, Austria. Association for Computational Linguistics.
- Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers in Artificial Intelligence*, 3(August):1–24.
- OpenRouter Inc. 2023. [OpenRouter API Documentation](#).
- Yuliia Polikovska. 2024. 47% of citizens receive information about events in Ukraine from Telegram channels – survey by the “Rating” group’. *Detector Media*.
- Hamidreza Saffari, Mohammadamin Shafiei, Hezhao Zhang, Lasana T. Harris, and Nafise Sadat Moosavi. 2025. Beyond hate speech: NLP’s challenges and opportunities in uncovering dehumanizing language. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26965–26980, Suzhou, China. Association for Computational Linguistics.
- Sam Schechner. 2022. Telegram Thrives Amid Russia’s Media Crackdown. *The Wall Street Journal*.
- Shalom H. Schwartz and Naomi Struch. 1989. Values, Stereotypes, and Intergroup Antagonism. In Daniel Bar-Tal, Carl F. Graumann, Arie W. Kruglanski, and Wolfgang Stroebe, editors, *Stereotyping and Prejudice*, pages 151–167. Springer New York, New York, NY.
- TGStat. 2025. [Telegram Channels and Groups Catalogue](#).
- Surendrabikram Thapa, Farhan Ahmad Jafri, Kritesh Rauniyar, Mehwish Nasim, and Usman Naseem. 2024a. RUHate-MM: Identification of hate speech and targets using multimodal data from russia-ukraine crisis. In *Companion Proceedings of the ACM Web Conference 2024*, Wwww ’24, pages 1854–1863, New York, NY, USA. Association for Computing Machinery.
- Surendrabikram Thapa, Kritesh Rauniyar, Farhan Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoglu, and Usman Naseem. 2024b. Extended multimodal hate speech event detection during Russia-Ukraine crisis - shared task at CASE 2024. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text (CASE 2024)*, pages 221–228, St. Julians, Malta. Association for Computational Linguistics.
- Yariv Tsfati and Jonathan Cohen. 2005. The Influence of Presumed Media Influence on Democratic Legitimacy: The Case of Gaza Settlers. *Communication Research*, 32(6):794–821.
- Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.
- G. Tendayi Viki, Daniel Osgood, and Sabine Phillips. 2013. Dehumanization and self-reported proclivity to torture prisoners of war. *Journal of Experimental Social Psychology*, 49(3):325–328.
- Michael J. A. Wohl, Matthew J. Hornsey, and Shannon H. Bennett. 2012. Why group apologies succeed and fail: Intergroup forgiveness and the role of primary and secondary emotions. *Journal of Personality and Social Psychology*, 102(2):306–322.
- Kateryna Zarembo, Michèle Knodt, and Jannis Kachel. 2024. Smartphone resilience: ICT in Ukrainian civic response to the Russian full-scale invasion. *Media, War & Conflict*, pages 305–324.

A Codebook for Manual Classification

Dehumanisation is the act of perceiving or treating people as less than human. In its most explicit forms, dehumanisation manifests itself in comparing of the other to an animal, machine, any non-human creature or inanimate substance. However, it may also be present through other manifestations, such as denial of agency, deindividuation, extreme negative evaluation as described in the codebook below.

This study focuses on dehumanisation of outgroup defined in national or political terms. You will be given a set of posts by Russian and Ukrainian Telegram channels published in the time of Russian invasion of Ukraine. Read carefully and identify which manifestations of dehumanisation are present in each post based on the given instructions, for each manifestation you should enter either ‘0’ or ‘1’. In addition to the codebook, you are provided with a table of group-identifying dehumanising keywords and killing euphemisms⁷, indicative of dehumanisation and specifically developed for the study. You can start with checking whether these words are present, entering ‘1’ into the respective manifestations columns and proceed by reading the whole post to fill in the remaining columns.

Every post should be classified as either dehumanising (‘1’), non-dehumanising (‘0’) or marginally dehumanising (‘0.5’). Posts containing an explicit comparison of the outgroup to an animal, machine, non-human creature, or substance should automatically be classified as dehumanising (‘1’), without assessing other manifestations. For posts without such explicit comparisons, classification should be based on the presence and overall strength of other dehumanisation markers. Posts that create a strong overall impression of the outgroup as less than human should be classified as dehumanising (‘1’), while weaker or partial manifestations should be classified as marginally dehumanising (‘0.5’). Only posts that contain no indicators of dehumanisation should be classified as non-dehumanising (‘0’).

Please, note the following:

1. If a formal group representative is dehumanised, it implies marginal dehumanisation of the whole group. Please, consult the keywords table for such cases.
2. Russian state-affiliated sources tend to divide Ukrainians into “good pro-Russia” and other “dangerous and aggressive” Ukrainians (those who are in favour of independent Ukraine) and typically deindividuates them. This should be treated as deindividuation of Ukrainian outgroup.
3. The words that fairly describe clearly observed behaviour, such as ‘Bucha massacre’, ‘okkupanty’ and ‘zakhvatchiki’ cannot be treated as signs of extreme negative evaluation or denial or prosocial behaviour, despite ‘fashisty’, which is used by both sides as an euphemism.

⁷The complete set of dictionaries used in the classification is available in [Chernenko \(2026\)](#)

B Comparison of Different LLMs for Dehumanisation Detection

Table 4: Model performance on the initial annotated sample with bootstrapped 95% confidence intervals, ordered by macro-F1 score

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
Dictionary classification	0.640 ± 0.061	0.639 ± 0.050	0.647 ± 0.059	0.608 ± 0.061
Llama 3.3 70B Instruct	0.810 ± 0.052	0.804 ± 0.050	0.810 ± 0.054	0.807 ± 0.055
GPT 4o-mini 1.63B	0.838 ± 0.050	0.794 ± 0.058	0.814 ± 0.053	0.800 ± 0.056
GPT 4o-mini 30.6B	0.838 ± 0.050	0.793 ± 0.061	0.810 ± 0.054	0.797 ± 0.058
Gemma 2 27B	0.748 ± 0.056	0.726 ± 0.052	0.746 ± 0.060	0.723 ± 0.057
DeepSeek 3 19.8B	0.775 ± 0.056	0.717 ± 0.063	0.737 ± 0.055	0.716 ± 0.063
Phi 3 14B	0.599 ± 0.065	0.516 ± 0.094	0.533 ± 0.047	0.481 ± 0.057
Mistral 7B	0.486 ± 0.065	0.541 ± 0.052	0.512 ± 0.050	0.441 ± 0.061
Llama 3.3 8B	0.441 ± 0.065	0.526 ± 0.082	0.463 ± 0.042	0.374 ± 0.066
Command R 35B	0.356 ± 0.063	0.625 ± 0.069	0.463 ± 0.061	0.373 ± 0.063
Qwen 2 7B	0.455 ± 0.065	0.528 ± 0.110	0.458 ± 0.037	0.367 ± 0.060
Phi 3 3.8B	0.378 ± 0.063	0.347 ± 0.054	0.268 ± 0.049	0.291 ± 0.049
Gemma 2 9B	0.297 ± 0.061	0.657 ± 0.173	0.397 ± 0.033	0.245 ± 0.056

C Additional Dataset Characterisation



Figure 2: Top 100 BOW unigrams (top) and bigrams (bottom) across dehumanisation categories in Russian Telegram channels



Figure 3: Top 100 BOW unigrams (top) and bigrams (bottom) across dehumanisation categories in Ukrainian Telegram channels

We have already discussed the most representative unigrams in dehumanising posts in Section 4.1. Ukrainian dehumanising bigrams similarly include expressions such as “russhists shelled,” “russhists military hardware,” and “Russian orcs.” Russian dehumanising bigrams include derogatory references to Ukrainians and their leadership, such as “a gang of drug addicts,” “Ukrainian Nazis,” and “banderovtsy naziki.” They also contain expressions specifically associated with channels linked to Chechen units of the Russian National Guard and Ramzan Kadyrov, including “Akhmat Sila” and “My dear brother,” reflecting these channels’ prominent role in producing dehumanising content.

Marginally dehumanising posts from Russian channels are characterised by words such as “adversary” or “enemy”, referral to negatively described Ukrainian nationalism (“Ukrainian nationalists”, “nazists”) and terms that refer disrespectfully to the Ukrainian state and army, diminishing their agency (“militants”, as Russian media sometimes describes Ukrainian army, “nationalist battalions”, “Kiev regime”). Marginally dehumanising posts from Russian channels are characterised by references to Ukrainians as adversaries or enemies, negative descriptions of Ukrainian nationalism (e.g., “Ukrainian nationalists,” “Nazis”), and terms diminishing the legitimacy or agency of Ukrainian institutions, such as “militants,” “nationalist battalions,” and “Kiev regime.” Ukrainian marginally dehumanising posts similarly describe Russians as adversaries or enemies, associate them with fascism, and employ derogatory terms such as “Moscovia” and “Horde” (“orda”). Frequent bigrams also reproduce the phrase “Russian warship, go f*ck yourself,”

which became a symbol of Ukrainian resistance following the Snake Island confrontation at the beginning of the full-scale invasion.

Finally, non-dehumanising posts are dominated by neutral wartime and political vocabulary, including country names and terms such as “president,” “briefing,” and “announced.” Ukrainian bigrams frequently reflect formal news reporting, such as “briefing by the presidential advisor.” Russian non-dehumanising posts are additionally dominated by fragments of the legally required disclaimer attached to content produced by media actors designated in Russia as “foreign agents,” suggesting that these independent or oppositional media actors disproportionately contribute to the non-dehumanising category.