Computational Analysis of Dehumanization of Ukrainians on Russian Social Media

Kateryna Burovova
LetsData / Lviv, Ukraine
kate.burovova@gmail.com

Mariana Romanyshyn
Grammarly / Kyiv, Ukraine
mariana.scorp@gmail.com

Abstract

Dehumanization is a pernicious process of denying some or all attributes of humanness to the target group. It is frequently cited as a common hallmark of incitement to commit genocide. The international security landscape has seen a dramatic shift following the 2022 Russian invasion of Ukraine. This, coupled with recent developments in the conceptualization of dehumanization, necessitates the creation of new techniques for analyzing and detecting this extreme violence-related phenomenon on a large scale. Our project pioneers the development of a detection system for instances of dehumanization. To achieve this, we collected the entire posting history of the most popular bloggers on Russian Telegram and tested classical machine learning, deep learning, and zero-shot learning approaches to explore and detect the dehumanizing rhetoric. We found that the transformer-based method for entity extraction SpERT shows a promising result of F_1 = 0.85 for binary classification. The proposed methods can be built into the systems of anticipatory governance, contribute to the collection of evidence of genocidal intent in the Russian invasion of Ukraine, and pave the way for large-scale studies of dehumanizing language. This paper contains references to language that some readers may find offensive.

1 Introduction

Dehumanization has been frequently proposed as a mechanism that mitigates or eliminates moral concern about cruel behavior, thus playing a crucial role in war, genocide, and other forms of extreme violence (Bandura, 1999). Recent research (Mendelsohn et al., 2020; Markowitz and Slovic, 2020; Magnusson et al., 2021) focuses on more subtle forms of dehumanization; those are considered both a precursor and a consequence of discrimination, violence, and other forms of day-to-day abuse outside of the context of armed conflicts. This shift in focus resulted in a simultaneously more nuanced and broad definition of dehumanization, inviting new approaches to operationalization. Multiple investigations (Diamond et al., 2022; Hook et al., 2023) have been conducted regarding the 2022 Russian invasion of Ukraine to determine if the Russian Federation is responsible for violating the Genocide Convention[1]. Central to these inquiries is the role of dehumanizing rhetoric in the direct and public encouragement of genocide. According to these reports, Russian officials and State media repeatedly described Ukrainians as subhuman ("bestial," "zombified"), contaminated or sick ("filth," "disorder"), or existential threats and the epitome of evil ("Hitler youth," "Third Reich," "Nazi"), rendering them legitimate or necessary targets for destruction. Hence, detecting the dehumanizing rhetoric at scale within this particular context can provide comprehensive evidence for further inquiries, as well as empirically support or challenge the assumptions of existing dehumanization frameworks.

2 Background

The concept of dehumanization has developed through cross-disciplinary conversations, integrating perspectives from various fields such as philosophy, psychology, sociology, and more. In the field of social psychology, Bandura (1999) investigated how people can psychologically detach themselves from others, viewing them as less than human, which can result in violence, bias, and discrimination. In sociology and critical theory, dehumanization was first scrutinized in relation to power dynamics, social disparities, and oppressive structures. Academics and thinkers such as Fanon (1967), Arendt (1963), and Bauman (1989) examined how dehumanization contributes to the marginalization, subjugation, and violence in sce-

[1] https://www.un.org/en/genocideprevention/genocide-convention.shtml

Proceedings of LaTeCH-CLfL 2024, pages 28–39
March 22, 2024 ©2024 Association for Computational Linguistics

narios of colonialism, totalitarianism, and genocide.

Dehumanization is often seen as a key stage leading to genocide[2]. However, Haslam (2019) argues that dehumanization is not just a precursor but is intertwined throughout the entire genocidal process. Our research shows that the temporal change of dehumanizing rhetoric on Russian Telegram conforms with this view.

3 Definition and Operationalization

Dehumanization is commonly defined as the denial of humanness to others (Haslam, 2006). Currently accepted frameworks mainly differ in understanding of "humanness" and of the ways in which this denial is taking place.

Kelman (1973) defines dehumanization as denying an individual both "identity" and "community"; Opotow (1990) introduces "moral exclusion" as an extension. Bandura (1999, 2002) argues that dehumanization relaxes moral self-sanctions and prevents self-condemnation. Harris and Fiske (2006, 2011) relate dehumanization to mental-state attribution, suggesting that dehumanized groups are perceived as having fewer mental states. This aligns with the mind perception theory by Gray et al. (2007), which categorizes perceptions into two dimensions: agency and experience.

The integrative review on dehumanization by Haslam (2006) proposes two distinct senses of humanness that can be denied in order to dehumanize persons or groups: human uniqueness **(UH)** and human nature **(NH)**. According to Haslam, the line dividing people from the related category of animals is defined by traits that are "uniquely human" (UH). Refined emotions, higher-order cognition, and language can all be considered uniquely human. Human-nature attributes (NH) are those that characterize humans in general and include emotional responsiveness, interpersonal warmth, openness, and emotional depth. Building on that, Haslam (2006) proposed two corresponding types of dehumanization: animalistic dehumanization, in which people or groups are thought to have fewer characteristics that make them uniquely human (and are perceived as vermin, animals, or disease), and mechanistic dehumanization, in which people or groups are thought to have fewer characteristics that describe people in general (and are perceived as automata or objects).

Li (2014) proposed the mixed model of dehumanization to address existing variability in the literature on dehumanization. This model is informed by framework by Haslam (2006); it contains four quadrants, formed by the level of Human Nature and Human Uniqueness attribution. We found this framework consistent with the most recent empirical evidence found in historical documents (Landry et al., 2022).

Genocide researchers highlight the limitations of using dehumanization as an early warning sign for atrocities. Neilsen (2015) introduced *toxification* as a more precise indicator. This concept goes beyond viewing victims as merely non-human and suggests that perpetrators see eradicating victims as essential for their survival, for two main reasons: victims are "toxic to the ideal" (threatening beliefs) or "toxic to the self" (posing harm).

Drawing from from Li (2014), Haslam (2006), and Neilsen (2015), we define *dehumanization* as the representation of the target group as existentially threatening and/or morally deficient by blatantly or subtly manipulating the features of its human uniqueness (including relevant elements in agency and competence) and/or human nature (including relevant elements in experience and warmth). Figure 1 summarizes all types with corresponding metaphors.

We chose the representation of Ukrainians in Russian Telegram as the target of our research. Below are some common blatantly dehumanizing metaphors used towards Ukrainians broken into types of dehumanization. Of type ↓ UH ↑ NH: укропитеки[3], свинорез[4], бандерлоги[5]; of type ↓ UH ↓ NH: расходный материал[6], майданутые[7], горящее сало[8], and of type ↑ UH

[2] http://genocidewatch.net/genocide-2/8-stages-of-genocide/

[3] [ukropiteki] — a derogatory term, combining "ukro" for "Ukrainian" and "piteki," which refers to early hominids

[4] [svinorez] — "pig slaughter," implying that Ukrainians are similar to pigs

[5] [banderlogi] — a play on Kipling's monkeys "Bandar-log" and "Bandera" (Ukrainian nationalist and the leader of the Ukrainian Insurgent Army)

[6] [raskhodnyy material] — "expendable material"

[7] [maidanutye] — the term is derived from "Maidan," the center of the Euromaidan protests in 2013-2014. The ending "-nutye" is common in Russian slang words meaning "crazy" or "nuts". Thus, the word can be translated as mentally ill with Maidan.

[8] [goryashcheye salo] — "burning lard," used to refer to Ukrainian people dying at the battlefield

↓ NH: укронацики[9], сатанисты[10], шайтаны[11]. ↑ UH ↑ NH means the absence of dehumanization.

In the case of укропитеки[3] , by being likened to pre-humans, Ukrainians are shown as lacking competence (dehumanized along the UH axis). Desires or experience are not denied, so NH axis position is unaffected; thus, we assign the label ↓ UH ↑ NH. The укронацики[9] metaphor demonizes Ukrainians and shapes an image of the epitome of evil, exaggerating competence and agency but reducing perceived warmth, affect, and shared human experience; thus we assign the label ↑ UH ↓ NH.

We treat dehumanization signals as additive. Therefore, compound words like Свинорейх[12], which consist of dehumanizing metaphors of ↓ UH ↑ NH and ↑ UH ↓ NH, are considered ↓ UH ↓ NH.

Table A.2 outlines a detailed description.

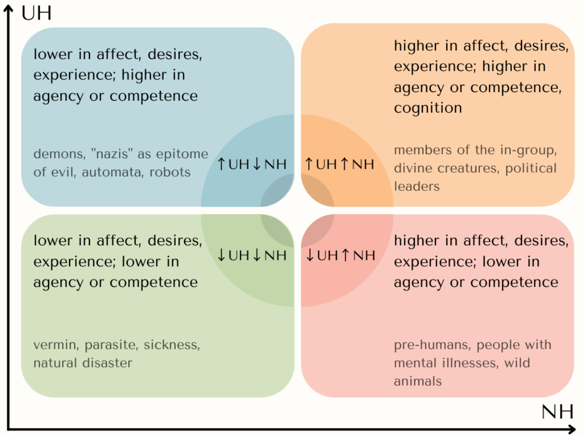
UH
lower in affect, desires, experience; higher in agency or competence
demons, "nazis" as epitome of evil, automata, robots
↑UH↓NH
↑UH↑NH
higher in affect, desires, experience; higher in agency or competence, cognition
members of the in-group, divine creatures, political leaders
lower in affect, desires, experience; lower in agency or competence
↓UH↓NH
↓UH↑NH
higher in affect, desires, experience; lower in agency or competence
vermin, parasite, sickness, natural disaster
pre-humans, people with mental illnesses, wild animals
NH

Figure 1: Four quadrants of dehumanization. ↑ UH ↑ NH represents the absence of dehumanization. The other three quadrants represent three types of dehumanization.

4 Related Computational Work

To the best of our knowledge, the first and only computational analysis framework focusing on dehumanization was proposed by Mendelsohn et al. (2020). It was applied to the analysis of discussions of LGBTQ people in the New York Times from 1986 to 2015. The authors identified linguistic correlates of salient components of dehumanization (negative evaluation, denial of agency, and metaphors of moral disgust and vermin) and then analyzed linguistic variation and change in discourses surrounding the chosen marginalized group.

In the study on dehumanization toward immigrants, Markowitz and Slovic (2020) attempted to evaluate the psychology of dehumanizers through language patterns, hypothesizing that three language dimensions reflect those who tend to dehumanize immigrants: (i) prevalence of impersonal pronouns, (ii) use of power words (e.g., "pitiful," "victim," "weak"), and (iii) emotion terms, evaluated through the affect category in LIWC (Pennebaker et al., 2015).

The study of Card et al. (2022) identifies dehumanizing metaphors by measuring the likelihood of a word denoting foreigners being related to a number of well-known dehumanizing metaphors (like "animals" and "cargo") in immigration-related sentences. The approach is best suited for the research setting where exhaustive lists of the considered metaphors are available.

Work by Magnusson et al. (2021) is informed by Bandura (1999); it presents a knowledge graph schema, dataset, and transformer-based NLP model SpERT to identify and represent indicators of moral disengagement and dehumanization in text. They define the multi-attribute knowledge graph extraction task as predicting the set of entities, the set of relations over entities, and the set of attributes over entities in a given text span. Among other indicators, they detect dehumanization based on the cumulative semantics of these attributes, entities, and relationships.

In the study of Nazi propaganda documents, Landry et al. (2022) analyzed the prevalence of agency and experience mental state terms used when referring to Jews in Nazi Germany, building on the moral disengagement theory by Bandura (1999) and mind perception theory introduced by Gray et al. (2007).

5 Research Setting

Existing solutions by Mendelsohn et al. (2020); Magnusson et al. (2021) have not yet been tested on cross-domain tasks. Computational analysis of dehumanization in the context of extreme violence (Landry et al., 2022) so far relied only on dictionary and lexicon-based approaches, and few dehumanization frameworks have been tested. Moreover, state-of-the-art large language models (LLMs) showing promising results across diverse

[9] [ukronatsiki] — a derogatory term which is a portmanteau of Ukrainian and Nazi

[10] [satanisty] — "satanists"

[11] [shaytany] — derived from the word "Shaytan" of Arabic origin, which means "devil" or "demon"

[12] [svinoreikh] — "Pig Reich" referencing the Nazi regime

domains were not yet tested for this task.

5.1 Methodology

Our approach is grounded in the definition of dehumanization proposed in Section 3. We narrow down the scope of the dehumanization to the context of extreme violence (hence, we consider only negative valence) but include both the subtle form expressed via metaphors and stylistic devices and the blatant form (e.g., directly likening the target group to inanimate objects).

We frame our primary task as the binary classification at the sentence level — detecting sentences containing at least one instance of dehumanization of Ukrainians. Building on the developed binary classification system, we work on a supplementary multi-class classification system that receives a sentence in Russian and classifies it by type according to the chosen dehumanization framework. We then investigate the explanatory potential of the chosen dehumanization framework.

Approach to Solution We start with collecting the data from selected social media sources. We proceed with the annotation project to obtain training data of the required granularity and format. We begin experimentation with the classical machine learning models to establish baseline performance and leverage their interpretability.

We experiment with enhancements like augmentation and feature engineering to improve performance. We then proceed with the deep learning approach by testing the SpERT model, applied for computational analysis of dehumanization for similar tasks by Magnusson et al. (2021). Next, we test the zero-shot learning approach using OpenAI (2022) GPT-3.5 Turbo[13]. We conclude experiments for the binary classification of dehumanization at the sentence level by comparing the results in the same setting. Next, we use the best model to explore the evolution of dehumanizing rhetoric within the timeframe of our dataset to test the explanatory potential of existing dehumanization frameworks.

For **quantitative evaluation** we rely on F_1, precision, and recall as our evaluation metrics. We adhere to an 80/20 train/test split with five folds for cross-validation.

Magnusson et al. (2021) reported micro-averaged F_1 = 50.12, precision 51.30, and recall of 51.29 for dehumanization relation for the SpERT model trained to extract signs of moral disengagement. We cannot treat these results as state-of-the-art and report these values purely as a reference, given that our results can not be directly compared due to the different contexts, underlying entity and relation schemes, and annotation approaches. Comparison with the commonly used methods is also impossible since we are pioneering binary dehumanization classification.

To further investigate the performance of the best-performing models, we perform a **qualitative error analysis** to identify the patterns in dehumanization not adequately captured by our models.

6 Dataset

For our analysis, we chose a group of 299 most popular political and news Telegram channels[14] (based on the ratings in the largest Telegram channels and groups catalog TGStat[15]) and collected their entire posting history spanning from 22 September 2015 to 25 November 2022, yielding 6.8M posts (23.91M sentence-level samples).

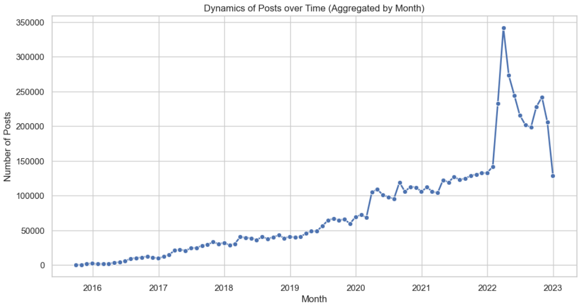

Figure 2: Dynamics of posts in the initial unlabeled dataset over time. Y-axis shows the absolute number of posts, and X-axis represents the time scale.

Several advantages arise from using social media as a data source for our task. Social media platforms provide (i) a dataset encompassing many different perspectives; (ii) an authentic snapshot of people's attitudes and behaviors; (iii) better accessibility with the APIs for ethical data collection.

While choosing a particular platform, we considered algorithmic contamination and the sanctions against Russia, due to which big social media companies have been limiting their functionalities[16].

Telegram is beneficial for our task due to (i) broad geographical scope and growth; (ii) being

[13] https://platform.openai.com/docs/models

[14] Unidirectional messaging platform where administrators can post exclusively.

[15] https://tgstat.com/

[16] https://www.npr.org/2022/02/26/1083291122/russia-ukraine-facebook-google-youtube-twitter

a primary messenger for most Russians (Newman et al., 2022); (iii) state role in decentralized content discovery: the Russian state uses the centralized endorsement by influencers, further amplifying their messages (Vavryk, 2022).

6.1 Data Annotation

To obtain training datasets we undertook the Annotation Project consisting of two sub-projects. First, we compiled a dataset of sentences and crowdsourced annotation with binary labels for the presence of dehumanization using Labelbox (2023). Then for the positive class sentences, we annotated dehumanizing spans by type of dehumanization.

We crowdsourced labels for this project from Ukrainian volunteers from the Ukrainian NLP community. We made this decision after carefully considering the alternatives: inviting Russian citizens (which would also introduce bias but would be much more difficult to set up) or inviting Russian-speaking annotators from other countries (who do not possess the needed level of context immersion). Since dehumanization is often expressed through the literary devices rooted in a particular culture, we clearly articulated how to handle the popular ambiguities in the annotation guidelines.

We evaluate the labels with Cohen's Kappa (Cohen, 1960), developed to account for the possibility that annotators guess on at least some variables due to uncertainty. The original annotation guidelines in Ukrainian for all sub-projects can be accessed via our GitHub repository[17]. All questions included the option to refuse answering if unsure, and all workers were clearly warned about the highly offensive content.

6.1.1 Part I of Annotation Project

The annotation schema for Part I of the annotation project (AP1) included three questions (listed here in translation from Ukrainian):

Q1 Does this sentence contain any mentions of Ukraine or Ukrainians?

Q2 Are there any comparisons that reduce Ukrainians to inanimate objects or individuals devoid of their distinctive human characteristics?

Q3 Is there an emotional evaluation of Ukrainians present in the text, and of what kind?

The full dataset encompasses various writing styles and spans years of posting history; thus, we expected the signals of dehumanization to be sparse. AP1 included two preselection phases. We started with a semi-manual random sentence sampling across the entire timeline and author set. To reduce the potential bias that this approach may impose on our training data, we finetuned transformer models for Russian on tasks of sentiment classification[18] and detection of mention of Ukrainians[19] using the Q1 and Q3 answers from the previous step. We then randomly sampled from the sentences from the full dataset classified by these models as containing a negative sentiment and a mention of Ukrainians.

Nine volunteers worked on AP1, annotating 4,111 samples in total. 39.28% of samples are positive dehumanization class, 20% of all samples were annotated by two workers. The overall inter-annotator agreement (IAA) is calculated as the pairwise mean. The labels for AP1 sentence-level classification are of high quality with Cohen's Kappa coefficients equal to 0.85 and 0.97 for dehumanization labels for the two preselection phases and the average of 0.90 and 0.92 for Q1 and Q3 respectively. Figure A.4 shows the distribution of classes for AP1.

6.1.2 Part II of Annotation Project

Our goal for Part II of the Annotation Project (AP2) was twofold: (i) to facilitate experimentation with entity classifiers, we need to obtain a dataset with spans labeled in the CoNLL04 format (Carreras and Màrquez, 2004); (ii) to track the evolution of dehumanizing rhetoric over time, we need to separate dehumanizing spans of different types. We used positive dehumanization class sentences from AP1 as the dataset for AP2 span annotation. The task was to identify spans of text that are dehumanizing towards Ukrainians and assign the correct dehumanization type to each span according to the three dehumanization quadrants of Figure 1. The guidelines contain detailed explanations for each type of dehumanization, provide examples, and cover instructions for edge cases, such as spans that combine multiple dehumanization types. These guidelines can be accessed through our GitHub

[17] https://github.com/kateburovova/dehumanization/tree/mainbranch/docs/annotation_guidelines

[18] https://huggingface.co/blanchefort/rubert-base-cased-sentiment

[19] https://huggingface.co/DeepPavlov/rubert-base-cased

repository[20]. A total of 478 sentences were annotated by one annotator. The majority class is ↑ UH ↓ NH. Figure A.5 shows the distribution of classes for AP2.

7 Experiments

7.1 Enhancements

In the initial phase of our research, we explored strategies to enhance the training by (1) extracting features with potentially stronger dehumanization signals and (2) generating synthetic training samples to reduce overfitting.

For the former, we drew from Mendelsohn et al. (2020) collocation extraction approach to extract four collocation types using spaCy[21]. We experimented with adding as features (i) verb-object or verb-adjunct collocations (ii) subject-verb collocations (iii) noun phrases where a noun is modified by another nominal element (iv) adjective-noun collocations.

For data augmentation, we employed an oversampling technique where we collected common non-dehumanizing mentions of Ukrainians or Ukraine and randomly replaced them in the data, thus generating new examples and reducing the reliance of the models on the context.

7.2 Classical Machine Learning

7.2.1 Logistic Regression

We started with Logistic Regression (LR) as the baseline classifier; the text was vectorized using the TF-IDF method. For this task, we used the dataset annotated during the AP1.

We experimented with clean lowercase (but not lemmatized) text, and then added lemmas and collocations as features. For each feature set, we performed grid search with cross-validation over the set of values for the parameters C (regularization strength) and penalty type (L1, L2). GridSearch for LR with lemmas and collocations as features produced the best result of F_1 = 0.78.

7.2.2 SVM

For experiments with SVM, we extended the same feature engineering approach. We performed a grid search over the regularization parameter C, which determines the balance between the misclassification of training examples and the simplicity of the decision surface. In all cases, the grid search relies on the best F_1 on the full test set as a selection criterion for each feature column. The best F_1 = 0.80 was attained with C=100 and linear kernel with enhancements; see the results in Table 1.

7.3 Deep Learning Approach

For experimentation with transformer models, we chose SpERT by Eberts and Ulges (2020), the attention model for span-based joint entity and relation extraction which performs the reasoning on BERT embeddings. For SpERT training, we use AP2 dataset. SpERT draws negative samples from the same sentence in a single BERT pass (Eberts and Ulges, 2020), so no additional negative samples were needed for training. In terms of classification, SpERT's entity recognition is a multiclass problem, where each identified span is associated with one of a predefined set of entity labels (but the spans can overlap partially or fully). To use SpERT in our setting of binary classification, we use a mapping function. Let $f(s)$ be the binary classification function, and $E(s)$ be the set of entities identified in a sentence s by the SpERT model. The function $f(s)$ maps to 1 if any entities are found in a sentence (i.e., if $E(s)$ is not empty), and 0 otherwise. This can be expressed as:

$$f(s) = \{ 1 , if E(s) \neq \emptyset, 0, if E(s) = \emptyset.$$

We followed the authors' recommendations on hyperparameter tuning, provided in their GitHub repository[22]. We report the SpERT model's performance in two contexts: multiclass classification over text spans produced best F_{1micro} = 0.80 and F_{1macro} = 0.81 and in binary setting F_1 = 0.85 as shown in Table 1.

7.4 Zero-Shot Learning

For experiments with zero-shot learning, we used ChatGPT gpt-3.5-turbo developed by OpenAI (2022) through the chat completions API endpoint. For this task, we used the dataset annotated during the AP1. Our approach was defined by the recommendations supplied by the OpenAI team[23] as well as by the empirical evidence shared within the developers community.

We evaluated 80 different prompt combinations for the GPT-3.5 Turbo agent, varying across three

[20] https://github.com/kateburovova/dehumanization/blob/mainbranch/docs/annotation_guidelines/An_part_II.pdf

[21] https://spacy.io/models/ru#ru_core_news_md

[22] https://github.com/markus-eberts/spert

[23] https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/

main components: definition of dehumanization (ranging from no definition to detailed guidelines in English or Ukrainian), the agent's role (from no specified role to acting as a social scientist, psychologist, or NLP researcher), and the approach to thinking process decomposition (from no specific instructions to step-by-step analysis strategies). The output formatting instructions for the agent remained consistent across all variations.

We found that the perspective of the social scientist induced the desired behavior, as well as the additional step of extracting the dehumanizing metaphors, if there are any, producing the best $F_1 = 0.82$.

7.5 Results and Discussion

Model	F_1	Precision	Recall
Logistic Regression	0.75	0.79	0.72
Logistic Regression with enhancements	0.78	0.82	0.75
SVM	0.75	0.79	0.72
SVM with enhancements	0.80	0.87	0.74
SpERT via a mapping function $f(s)$	**0.85**	**0.86**	**0.85**
GPT-3.5 Turbo	0.82	0.86	0.82

Table 1: Results for all models in the binary setting on the test set.

Through the employment of LR as the baseline classifier, we confirmed lemmatization as a critical pre-processing step and subject-verb structures as key features. During experimentation with LR, we observed that the models detect blatant dehumanization (with dehumanization contained directly in group labels) better than subtle dehumanization (usually expressed via metaphors and stylistic devices). This encouraged us to split the test set into blatant and subtle dehumanization subsets based on the presence of dehumanization in group labels' spans and additionally test the models' performance on them.

The implementation of the SVM model revealed an improved performance over the LR baseline. The SVM model demonstrated a propensity towards precision, aiming to minimize false positives. Augmenting the dataset enhanced performance on the samples with subtle dehumanization. Using SpERT in the binary setting, we reached SOTA performance for the binary dehumanization classification task at the sentence level. We concluded our experimentation with GPT-3.5 Turbo. By testing various combinations of context prompts we observed that the perspective of a social scientist and additional steps for extracting dehumanizing metaphors produced the most desirable results.

We can report that SpERT and GPT-3.5 Turbo showed significantly better results for the detection of subtle dehumanization than Logistic Regression or SVM. While results for the blatant dehumanization subset were comparable, best SVM model attained only $F_1 = 0.51$ on the subtle dehumanization subset, GPT-3.5 Turbo showed significant improvement with $F_1 = 0.65$ and best SpERT model produced the result of $F_1 = 0.82$. Precision and recall are much better balanced for SpERT and GPT-3.5 Turbo than for LR or SVM as well.

We observed higher average performance of the GPT-3.5 Turbo model, when prompted with the original Ukrainian annotation instructions on the test subset featuring dehumanization in group labels. This implies that GPT-3.5 Turbo utilized the examples in the original text to effectively match phrases closely related to those examples.

8 Temporal Dynamics of Dehumanization

To investigate how types of dehumanization evolved over time, we used SpERT in the multiclass setting to detect instances of dehumanization throughout the initially collected Telegram dataset.

We calculated the representative sample size for each period to assess the dynamics of dehumanization by type for 95% confidence interval with 1% margin of error, accounting for the Bessel's correction. All required samples sizes were below 300 posts, we chose 1,000 posts per time period as a reasonable sample size.

Figure 3 shows the dynamics for dehumanization by type. We added two notable events to the plot: the vertical blue line signifies the date of publication of the essay "On the Historical Unity of Russians and Ukrainians"[24], in which Putin publicly questions the legitimacy of Ukraine as a state, and the vertical red line shows the start of the 2022 Russian invasion of Ukraine. We observe that the dehumanization rhetoric, manipulating both dimensions of humanity (LOW_UN_LOW_NH) is the only type following a stable growth pattern over time. This is the type of dehumanization that is con-

[24] http://en.kremlin.ru/events/president/news/66181

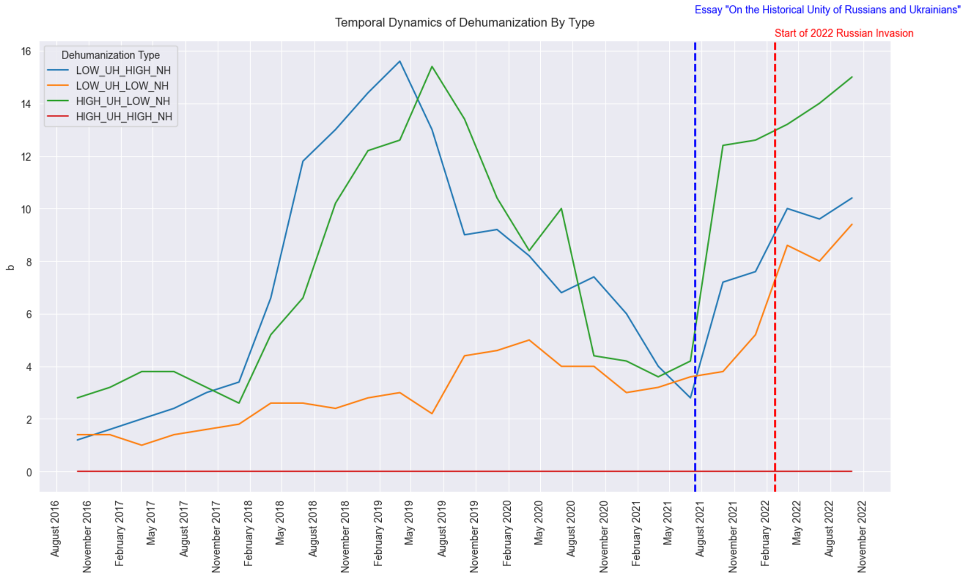

Figure 3: Dynamics of dehumanization. The X-axis represents a time scale, with data points aggregated over 3-month intervals with granularity of 1 month for random samples. The Y-axis depicts the mean number of positives.

sidered indicative of extreme violence risks and includes disgust-driven dehumanization and objectification. This dynamic does not appear to be defined by swift changes in political background or key policy-maker's decisions. The hyper-humanization (HIGH_UN_HIGH_NH) is not present due to its absence in the training data.

The two types of dehumanization that manipulate one of the two dimensions of dehumanization (HIGH_UN_LOW_NH and LOW_UN_HIGH_NH) demonstrate complex patterns. Their frequency starts to increase around 2017, and peaks in 2019 around the time of Ukrainian presidential elections[25], reaching the lowest point in 2021 by the time of the first wave of Russia's amassing troops at Ukraine's borders[26]. The rapid changes in these dehumanization types suggest that the dissemination of the imagery they supply may be orchestrated, or they are highly sensitive to the shifts in political reality. Figure 3 shows that these types start to increase not long before the 2022 invasion and drop at its start, confirming the idea of the preparatory role of dehumanizing rhetoric in sanctioning genocide. Notably, the dehumanization signals do not return to the pre-invasion levels with time, suggesting that this phenomenon cannot be localized as only the precursory stage in extreme violence. We observe that different types of dehumanization are evolving at different pace, suggesting that each fulfills a specific role.

9 Conclusion

In this research, we have delved into advanced techniques for dehumanization detection in the backdrop of extreme violence, culminating in the development of the first-ever dataset in the Russian language annotated at both sentence and span levels and a SpERT-based dehumanization detection model showing $F_1 = 0.85$. Leveraging our state-of-the-art model, our work offers a clear window into the temporal dynamics of dehumanizing rhetoric both before and during the 2022 Russian invasion of Ukraine, setting a precedent in the field. This system holds potential for integration into systems aimed at predicting and preventing extreme violence and creates a foundation for further research of computational analysis of dehumanization. Both best SpERT model and dataset are available for non-commercial use upon reasonable request and following the intended use; they have not been made publicly accessible to prevent potential malicious use.

10 Ethics and Limitations

We lack the instruments to compile a dataset representative of the general structure of the Russian population; instead, we focus on the most influen-

[25] https://www.bbc.com/news/world-europe-48007487

[26] https://www.iiss.org/online-analysis/online-analysis/2021/12/why-is-russia-amassing-troops-at-its-border-with-ukraine

tial media figures to infer the state of public thought from the speech patterns they are spreading. We do not intend to draw causal inferences between the magnitude of the type of dehumanization signal and the severity of violence toward Ukrainians. Instead, we seek to examine which dehumanizing perceptions are implicated in harm and how the change in degree and components can evolve.

Our toolkit and operationalization techniques can be extended to detect dehumanization in languages other than Russian. However, adaptation to other languages and contexts would require accounting for their unique cultural and political landscapes.

The annotated dataset was ensured to contain no publicly identifiable information (PII) other than widely available media coverage.

References

Hannah Arendt. 1963. *Eichmann in Jerusalem: A Report on the Banality of Evil*. Viking Press.

Albert Bandura. 1999. Moral Disengagement in the Perpetration of Inhumanities. *Pers Soc Psychol Rev*, 3(3):193–209.

Albert Bandura. 2002. Selective Moral Disengagement in the Exercise of Moral Agency. *Journal of Moral Education*, 31(2):101–119.

Zygmunt Bauman. 1989. *Modernity and the Holocaust*. Cornell University Press.

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 89–97, Boston, Massachusetts, USA. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46. Place: US Publisher: Sage Publications.

Yonah Diamond, John Packer, Farrell Rosenberg, and Susan Benesch. 2022. An independent legal analysis of the russian federation's breaches of the genocide convention in ukraine and the duty to prevent.

Markus Eberts and Adrian Ulges. 2020. Span-based Joint Entity and Relation Extraction with Transformer Pre-training. *Santiago de Compostela*.

Frantz Fanon. 1967. *Black Skin, White Masks*. Grove Press.

Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6):878–902.

Heather M. Gray, Kurt Gray, and Daniel M. Wegner. 2007. Dimensions of Mind Perception. *Science*, 315(5812):619–619.

Lasana T. Harris and Susan T. Fiske. 2006. Dehumanizing the Lowest of the Low: Neuroimaging Responses to Extreme Out-Groups. *Psychol Sci*, 17(10):847–853.

Lasana T. Harris and Susan T. Fiske. 2011. Dehumanized Perception: A Psychological Means to Facilitate Atrocities, Torture, and Genocide? *Zeitschrift für Psychologie*, 219(3):175–181.

Nick Haslam. 2006. Dehumanization: An Integrative Review. *Pers Soc Psychol Rev*, 10(3):252–264.

Nick Haslam. 2019. The Many Roles of Dehumanization in Genocide. pages 119–138. Oxford University Press. Book Title: Confronting Humanity at its Worst.

Kristina Hook, John Packer, Farrell Rosenberg, and Susan Benesch. 2023. The russian federation's escalating commission of genocide in ukraine: A legal analysis.

Herbert G. Kelman. 1973. Violence without Moral Restraint: Reflections on the Dehumanization of Victims and Victimizers. *Journal of Social Issues*, 29(4):25–61.

Labelbox. 2023. Labelbox. [Online]. Available: https://labelbox.com/.

Alexander P. Landry, Ram I. Orr, and Kayla Mere. 2022. Dehumanization and mass violence: A study of mental state language in Nazi propaganda (1927–1945). *PLoS ONE*, 17(11):e0274957.

Mengyao Li. 2014. Towards a comprehensive taxonomy of dehumanization: Integrating two senses of humanness, mind perception theory, and stereotype content model.

Ian H. Magnusson, S. Schmer-Galunder, Ruta Wheelock, Jeremy Gottlieb, Pooja Patel, and Christopher Miller. 2021. Toward Transformer-Based NLP for Extracting Psychosocial Indicators of Moral Disengagement. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 43*.

David M. Markowitz and Paul Slovic. 2020. Social, psychological, and demographic characteristics of dehumanization toward immigrants. *Proc. Natl. Acad. Sci. U.S.A.*, 117(17):9260–9269.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A Framework for the Computational Linguistic Analysis of Dehumanization. *Front. Artif. Intell.*, 3:55.

Rhiannon Neilsen. 2015. 'Toxification' as a More Precise Early Warning Sign for Genocide Than Dehumanization? An Emerging Research Agenda. In *Genocide Studies and Prevention*, volume 9, pages 83–95. ISSN: 1911-0359, 1911-9933 Issue: 1 Journal Abbreviation: GSP.

Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, and Rasmus Kleis Nielsen. 2022. Digital news report 2022. Technical report, Reuters Institute for the Study of Journalism.

OpenAI. 2022. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Susan Opotow. 1990. Moral Exclusion and Injustice: An Introduction. *Journal of Social Issues*, 46(1):1–20.

James Pennebaker, Roger Booth, Ryan Boyd, and Martha Francis. 2015. *Linguistic Inquiry and Word Count: LIWC2015*.

Petro Vavryk. 2022. Mapping Growth of the Russian Domestic Propaganda Apparatus on Telegram. *Challenges to national defence in contemporary geopolitical situation*, 2022(1):227–231.

A Appendix

Label	Type
↓ UH ↓ NH	Low in Human Nature and Low in Human Uniqueness quadrant corresponds to disgust-driven dehumanization and objectification. Welfare recipients, drug users, and the homeless are among the social groups most at risk from this severe dehumanization (Fiske et al., 2002). These groups, which are perceived as detached and incapable, arouse intensely unpleasant emotions like disgust and hatred, which in turn predict both active harm (harassment) and passive harm (neglecting) behavioral patterns.
↑ UH ↓ NH	Low in Human Nature and High in Human Uniqueness quadrant corresponds to the mechanistic dehumanization, and members of groups dehumanized in this manner are often perceived as cold, rigid, passive, and yet highly competent (e.g., technicians, businesspeople). Mechanistic dehumanization denotes a horizontal social comparison to unfamiliar individuals and elicits responses like indifference and alienation instead of contempt and denigration (Haslam, 2006), in contrast to animalistic dehumanization, which reflects a downward social comparison. Superhumanization and demonization fall into this category. Demonization, in particular, is a common technique in acts of extreme violence, in which the target is branded as evil and incapable of change [(Li, 2014)]. The roles of perpetrators and victims are flipped completely when violence victims are demonized. For instance, during the Holocaust, the persecutors saw themselves as heroes for ensuring the survival of a superior race while portraying Jews as evil criminals (Landry et al., 2022). By doing this, demonization not only excludes victims from moral consideration (Opotow, 1990), but it also establishes a moral mandate that labels victims as evil and calls for action to be taken against them.
↓ UH ↑ NH	High in Human Nature, Low in Human Uniqueness quadrant corresponds to animalistic dehumanization. When UH is thought to be absent, people are frequently negatively viewed as unintelligent, impolite, or lacking in self-control resembling non-human animals. However, the perceived high levels of NH are linked to a concurrently neutral or even favorable assessment of others as warm, emotional, and creative. This form of dehumanization treats dehumanized targets as unrefined animals without necessarily subjecting them to malicious prejudice and inhumane treatment. This perception is consistent with the paternalistic stereotype in the Stereotype Content Model (SCM), which appears predominantly in traditional portrayals of women, elderly, or the disabled (Fiske et al., 2002).
↑ UH ↑ NH	High in Human Nature and high in Human Uniqueness quadrant corresponds to humanization and superhumanization. Some people and groups are seen as fully human on both dimensions, which is the opposite of the extreme dehumanization with both Human Nature and Human Uniqueness denied. According to the SCM, ingroup members are typically viewed as both warm and competent, which is consistent with the idea of ingroup favoritism, or the tendency to favor the ingroup over the outgroup.

Table A.2: Detailed Description of Dehumanization Types

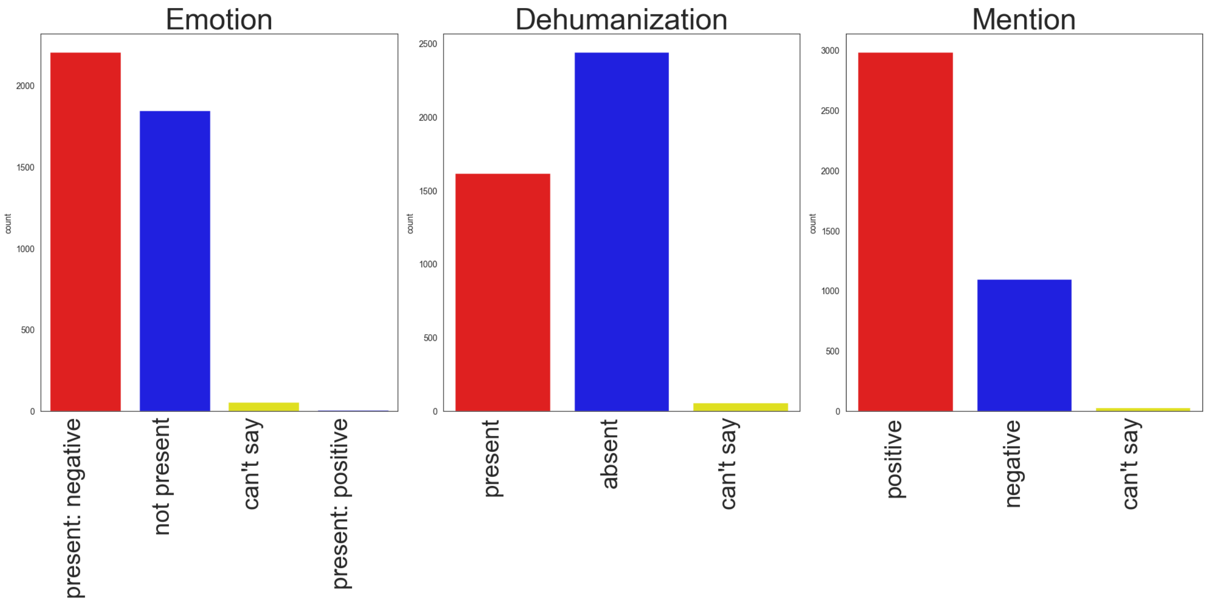

Figure A.4: The distribution of classes in the AP1.

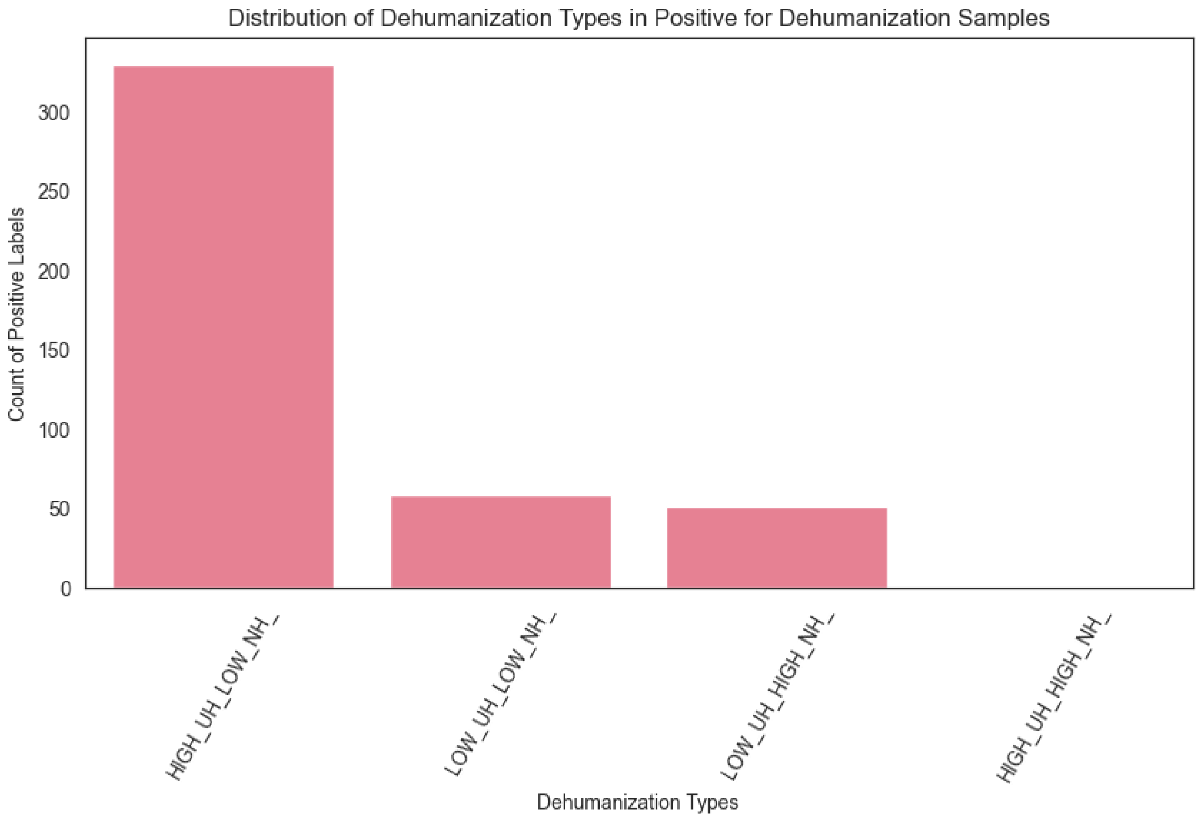

Figure A.5: The distribution of classes in the AP2.