CHARMemes: Collection of Harmful Runet Memes

Anonymous ACL submission

Abstract

Harmful memes pose a distinct challenge for content moderation: they compress social meaning into a visual template with short text, often relying on irony, intertextuality, and local cultural context. Current multimodal safety datasets either focus on narrow harm definitions or mainly English meme ecosystems, limiting progress on robust, culturally grounded detection. We introduce CHARMemes [1] a large-scale benchmark for harm-aware meme understanding in the Russian-oriented online ecosystem (Runet). The dataset contains 23,025 in-the-wild memes collected from a template-centric public source and spans 2007–2025, enabling temporal and cross-domain analyses. We propose a taxonomy of 11 harm categories grouped into three severity buckets, and build the dataset with a scalable pipeline combining automated labeling, targeted human review, and multi-stage deduplication.

We benchmark multiple vision–language approaches under binary, severity-level, and fine-grained settings, and evaluate robustness across time and datasets. Results indicate that binary harmfulness detection is comparatively strong, while fine-grained harm recognition remains difficult, especially under temporal shift. CHARMemes offers a realistic testbed for developing more robust harmful meme detectors.

Warning: this paper contains example data that may be offensive, harmful, or biased.

1 Introduction

In today's digital culture, memes are a pervasive form of everyday expression and public discourse. More broadly, a meme can be treated as a digital object-most often an image with a caption, but also a video, a catchphrase, or other reusable format-that is collectively created, transformed, and circulated online (Shifman, 2013). Memes operate as a communication genre (Wiggins, 2019), closely linked to participatory culture that encompasses both creative cultural play and more destructive forms of engagement such as trolling or harassment (Nagle, 2017; Tuters and Hagen, 2020). At the same time, these dynamics also enable explicitly harmful uses of memes in contemporary political and socio-cultural contexts.

Figure 1: The meme on the left was classified as Offensive (General Vulgarity), while the meme on the right was considered Safe.

Such harmful memes target individuals, communities, or social groups, reinforcing discrimination, spreading hate or misinformation, and exacerbating social divisions (Pramanick et al., 2021b; Kiela et al., 2020; Sabat et al., 2019). Their rapid circulation on social media amplifies their societal impact, making the detection and mitigation of harmful memes a crucial and challenging task for content moderation and machine learning research. Despite this growing attention, the research landscape is uneven across languages.

This unevenness is especially problematic because memes are inherently culture-indexed: their meaning is often carried by shared background knowledge, including local events, public figures, slang, and platform-specific conventions, rather than literal text alone. Consequently, the same template or caption can shift semantics across regions and languages, and models that perform well on English-centric benchmarks may fail when confronted with culturally grounded references. For instance, in Figure 1, the visual template of the

[1] A link to the dataset and code will be provided in the camera-ready version to preserve anonymity during review.

meme is the same, however, the semantics and texts shift the meaning and perception of the meme. Moreover, regional bias in memes is frequently implicit, encoded through stereotypes, in-group humor, and geopolitical or identity framing, which makes it difficult to detect using surface lexical cues alone. Current work on harmful memes, however, remains largely English-centric. Moreover, many existing studies and benchmarks are organized around time-bounded, high-salience events such as the COVID-19 pandemic (Pramanick et al., 2021b), election cycles (Pramanick et al., 2021a), or armed conflicts (Thapa et al., 2022) rather than modeling harmful memetic content as a persistent, evolving phenomenon. Existing datasets and models primarily focus on identifying harmful memes and their targets (Pramanick et al., 2021a; Sharma et al., 2023), detecting propaganda techniques in memes (Dimitrov et al., 2021), and performing fine-grained classification of phenomena such as racism, sexism, and antisemitism (Zia et al., 2021; Chandra et al., 2021). Beyond English, resources exist only for a small number of languages, for example Tamil troll memes (Suryawanshi et al., 2020b), Bengali (Hossain et al., 2022), and Chinese hateful memes (Lu et al., 2024), leaving many large non-English online spheres underexplored. A notable gap is the Russian-oriented online ecosystem, which constitutes a large and socially consequential non-English sphere. Russian is among the world's most widely spoken languages [2] and is also among the most common content languages on the web [3], and the broader Russian-oriented online ecosystem–often referred to as Runet–plays a central role in political and cultural discourse across the post-Soviet space.

We use Runet in this broader sense, encompassing not only Russian-language content but also other languages and communities in the post-Soviet online space. Yet, to the best of our knowledge, comparable resources for studying harmful memes in Runet remain largely absent.

To fill-in this research gap we:

- Introduce `CHARMemes` a large-scale benchmark of memes from the Runet ecosystem, designed to support harm-aware multimodal understanding over a long temporal span.
- Propose a fine-grained harm taxonomy with 11 categories organized into severity buckets, enabling more nuanced analysis than binary harmfulness detection.
- Provide strong baselines and robustness analyses, including fine-grained vs. binary evaluation and studies of temporal and cross-dataset generalization, highlighting current limitations and open challenges.

We also release our dataset and the code for data scraping, experiments, and analysis[4].

2 Related work

Memes in NLP and Vision–Language Research. In the NLP and vision–language literature, memes are widely used as a stress test for multimodal understanding because their meaning is often not literal: it emerges from the interaction of the image, overlaid text, template conventions, and shared cultural knowledge, frequently involving humor, sarcasm, or implicit references. To make this challenge measurable, prior work has mainly framed "meme understanding" as supervised benchmark tasks—e.g., predicting affect or humor in Memotion (Sharma et al., 2020), identifying multimodal social harms such as misogyny in SemEval MAMI (Fersini et al., 2022), or performing target-aware hateful meme classification in the Hateful Memes challenge (Kiela et al., 2021). Methodologically, progress has largely tracked advances in vision–language representation learning and pretraining, where transformer-based multimodal encoders and contrastive objectives improve the alignment between text and images (Li et al., 2019; Chen et al., 2019; Kim et al., 2021; Radford et al., 2021). Despite these improvements, meme understanding remains difficult because key context is frequently unstated and the same template can express different intents across communities and over time, which induces strong distribution shift and brittle generalization (Kiela et al., 2021).

Multimodal Hate and Harmful Meme Detection. A major line of work studies memes through the lens of *harm*: hate speech, offensiveness, and related safety risks that emerge only when image and text are interpreted jointly. Early influential benchmarks formalized this as supervised classification with explicit multimodal stress tests-for example, the *Hateful Memes* challenge constructs

[2] https://www.cia.gov/the-world-factbook

[3] `https://w3techs.com/technologies/overview/content_language` (accessed: Dec 2025)

[4] A link will be provided in the camera-ready version to preserve anonymity during review.

Domain	Definition	Category
High-severity harms	Most likely to involve coercion, non-consent, or acute physical/psychological danger (Krug et al., 2002; Seko and Lewis, 2018; Arendt et al., 2019).	Sexual Exploitation Violence Self-Harm
Mid-severity harms	Centers on demeaning or targeting people and can lead to real-world harms through stigmatization or coordinated abuse (Fortuna and Nunes, 2018; Rosa et al., 2019; Vogels, 2021).	Hate Speech Harassment
Contextual / platform-policy harms	May be harmful, misleading, or inappropriate depending on framing and audience, but is often less dangerous than high-severity harms (Scheuerman et al., 2021; Jiang et al., 2021).	Animal Cruelty Illegal Content Propaganda Offensive NSFW
Safe	Does not fall into any of the above harm categories.	Safe

Table 1: Harm taxonomy grouped into severity buckets with our new categories in orange.

"benign confounders" to reduce unimodal shortcuts and shows a large gap between human performance and strong model baselines, highlighting the difficulty of robust multimodal hate detection (Kiela et al., 2021). Beyond binary hate, several datasets broaden the notion of harm: *MultiOFF* targets offensive meme content in politically themed memes and compares early-fusion multimodal models against unimodal baselines (Suryawanshi et al., 2020a), while *HarMeme* introduces harmfulness detection together with identifying the *targeted social entities*, motivating target-aware modeling for harmful memes (Pramanick et al., 2021b,a). Shared-task settings such as SemEval MAMI further emphasize fine-grained harmful categories (e.g., types of misogyny) and have driven practical architectures that combine OCR text, visual encoders, and multimodal fusion (Fersini et al., 2022). Overall, progress has largely come from better multimodal fusion and pretraining, but performance remains brittle because meme harm is often implicit, culturally grounded, and sensitive to small changes in either modality (Kiela et al., 2021; Pramanick et al., 2021b).

3 Harmful Meme Categorization

As reviewed in Section 2, harmful meme benchmarks span heterogeneous label spaces, ranging from binary hate detection to task-specific taxonomies (e.g., targets, misogyny, propaganda), which makes results hard to compare and can underspecify in-the-wild phenomena. Since `CHARMemes` aims to evaluate harm in a large, culturally grounded Runet ecosystem, we need a unified categorization scheme that covers recurring harm types in real meme sharing, provides clear boundary rules for ambiguous cases, and enables evaluation at multiple granularities. We therefore build the harm taxonomy below and use it consistently for automated labeling, human adjudication, and as the label space for evaluation in Section 5.

We categorize harmful memes by adapting the typology of Sharma et al. (2022). Building on their broad families, we refine the scheme to better match contemporary in-the-wild meme usage and to support clearer analysis across platforms. Our taxonomy contains 11 labels (Table 1), which we additionally group into three severity buckets: high-severity, mid-severity, and contextual/platform-policy harms. *Safe* denotes memes that do not fall into any harm category.

Our differences from prior typologies are driven by two considerations: coverage and boundary clarity. We add three explicit categories that are frequent in real-world meme sharing, but are often implicit or merged into broader labels in earlier schemes: Illegal Content, Animal Cruelty, and NSFW. Treating these as first-class categories avoids collapsing heterogeneous phenomena into a residual bucket (e.g., "offensive") and supports more informative analyses of which harms are present and how often they occur. This choice is also motivated by the way contemporary platforms are increasingly used to distribute and normalize regulated or illicit material: for example, social media has become an important venue for drug advertising and dealing-related activity (Demant and Aagesen, 2022). Likewise, animal cruelty has emerged as a distinct, highly shareable (and sometimes monetized) genre of online content, with evidence of systematic exploitation of animals by content creators across platforms and

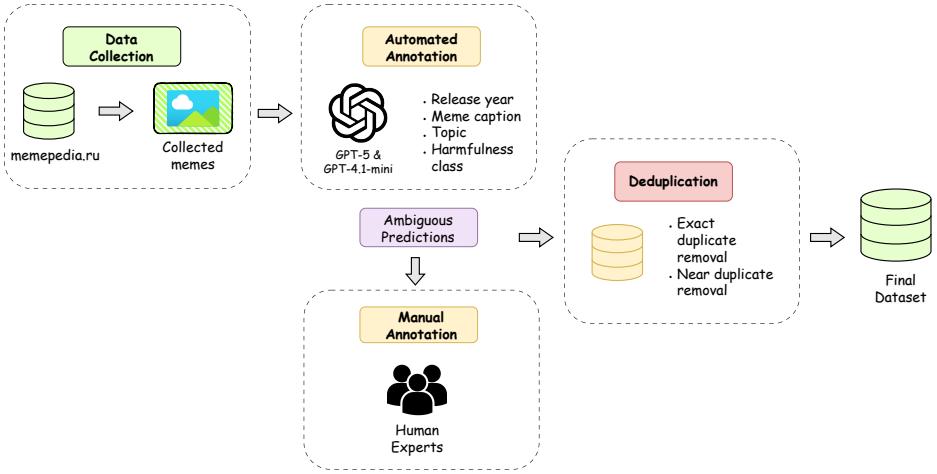


Figure 2: Dataset Creation Pipeline: we scrape memes from memepedia.ru, then we do automated annotation with state of the art commercial LLMs, ambigous samples are re-annotated by human annotators. Finally we remove duplicates from the dataset.

countries (Carvalho et al., 2023).

Because meme meaning is often indirect and context-dependent, category boundaries can be particularly ambiguous. We separate legal, consensual adult sexual or suggestive content (NSFW) from sexual content involving coercion (McGlynn and Rackley, 2017), non-consent, blackmail/trafficking, or minors (Wolak et al., 2018) (Sexual Exploitation). Similarly, we distinguish Hate Speech - attacks on protected groups - from Offensive content (crude or insensitive content without protected-group animus) (Davidson et al., 2017), and separate Harassment - targeting identifiable individuals, including doxing or brigading (Thomas et al., 2021) - from group-directed hostility, which falls under Hate Speech. Finally, we treat self-harm as a dedicated harm family, since suicide/self-injury and eating-disorder promotion differ in both manifestation and intervention needs (De Choudhury et al., 2016) from other forms of harmful content.

For added granularity, we map each top-level category to a set of representative subtypes (illustrative rather than exhaustive). Hate Speech may include racism/xenophobia, sexism, homophobia/transphobia, religious hate, ableism, and other forms of targeted hostility toward protected groups. Violence may include depictions or glorification of violence, as well as terrorism/extremism-related content. Illegal Content may include drugs and gambling, along with scams, fraud, and other illicit or regulated activities. Self-Harm may include suicide/self-injury content and the promotion or encouragement of eating disorders. NSFW may include nudity or sexually suggestive content and adult sexual humor. Full definitions, subtype descriptions, and representative examples for all categories are provided in the annotation guidelines in Appendix C.

4 Dataset

In this section, we describe the construction of CHARMemes, outlining how we collect items from real-world meme sources and organize them into a unified benchmark. We then summarize the resulting dataset contents and the basic processing/annotation steps needed to make it usable for downstream evaluation. The dataset creation pipeline is shown on Figure 2, and detailed in following sections. Dataset statistics are given in Appendix A.

4.1 Data Collection

Following Bates et al. (2025), we scraped Memepedia, a Russian meme database[5] that is structurally similar to Know Your Meme (KYM)[6]. Each Memepedia entry provides a template-level description - origin and context - and multiple "in-the-wild" usage examples sourced from web forums and social media. For each meme, we parsed the template description and all available usage examples.

From the template description, we extracted the meme's year of release using GPT-4.1-mini for cost efficiency. We manually verified this extraction on a random sample of 300 templates, confirming that

[5] https://memepedia.ru/
[6] https://knowyourmeme.com/

Dataset	# Samples	Time span	Is multilingual?	Fine-grained?	# Topics
Hateful Memes Challenge (Kiela et al., 2020)	10,000	2020	No	Yes	1
HarMeme (Pramanick et al., 2021a)	7,096	2020–2021	No	Yes	2
MultiOFF (Suryawanshi et al., 2020a)	743	2016	No	No	1
CHARMemes	**23,025**	**2007–2025**	Yes	**Yes**	**20+**

Table 2: Comparison of multimodal meme datasets.

the predicted year matches the year stated in the description.

We annotated each usage example with GPT-5 according to the taxonomy in Section 3, using the prompt in Appendix D. To improve reliability, we used a three-round re-annotation/self-consistency procedure similar to MemeMind (Gu et al., 2025). If GPT-5 produced the same label in all three rounds, we accepted it as final. Otherwise, we treated the example as anomalous and routed it to human annotation.

As additional metadata, we added meme caption's transcript, its language, detected with lingua[7], and a meme's topic label inferred with GPT-5 in an additional round of annotation (See Appendix J).

4.2 Human Annotation

While automated labeling covers the full corpus, we reserve human annotation for anomalous or ambiguous cases flagged by our self-consistency procedure. As a result, nearly 20% of the dataset were adjudicated by trained bilingual annotators under a shared set of guidelines. Our annotators are native Russian speakers with at least C1 English proficiency familiar with Runet meme culture (active social media users since at least 2013). Annotators were compensated on an hourly basis, with rates aligned to prevailing local wages in their country of residence[8]. We provided a labeling interface with clear annotation guidelines (see Appendix G) and ensured that each example was annotated independently by at least two annotators. When annotators agreed, we used the shared label; when they disagreed, they discussed the case and assigned a consensus label. In Appendix I we provide some examples of inconsistent GPT annotations and consensus human labels.

To validate GPT-5 labels, we randomly sampled 200 examples from the subset with three-round LLM agreement and asked annotators to label them as a trial task. We then measured inter-annotator agreement and annotator-LLM agreement (Appendix E). One annotator's labels deviated substantially from the rest; after adjudication and review, we excluded this annotator from further labeling due to numerous errors in annotation.

4.3 Deduplication

To ensure dataset quality, we implemented a three-stage deduplication pipeline for our meme images. First, we implemented MD5 hashing to remove identical images, which resulted in removal of 94 exact duplicates. Second, we applied perceptual hashing to detect near-identical images (Hamming distance ≤ 5), removing 206 samples. Third, we used multimodal similarity computed using ResNet-50 (He et al., 2016) for image embeddings and a sentence-transformer[9] (Reimers and Gurevych, 2020) for text embeddings. We chose these models as a lightweight, high-throughput embedding pipeline: ResNet-50 provides compact visual features that are well-suited to capturing mid-level pattern similarity (sufficient for our duplicate filtering), and multilingual-MiniLM-L12-v2 offers strong semantic representations at a small model size and low inference cost. Using these embeddings, we removed 412 samples with ≥ 0.95 similarity in both modalities. To further detect near-identical images, we classified borderline cases that had similarity scores between 0.85 and 0.95 (n = 1702) for manual review. During manual review, we further identified 633 duplicates and restored 30 samples that were flagged as duplicates by our pipeline. In total, our pipeline removed 1345 samples, resulting in 23,025 unique memes. Examples of such memes can be found in Appendix H.

4.4 Dataset Comparison

Table 2 summarizes how CHARMemes compares to prior multimodal meme datasets. Owing to careful curation, CHARMemes is substantially larger than existing benchmarks and spans a much longer time range, enabling analyses across multiple meme

[7] https://github.com/pemistahl/lingua-py

[8] Additional annotator details are provided in Appendix F.

[9] multilingual-MiniLM-L12-v2

"eras" rather than a single snapshot. Unlike most prior datasets, which are English-only and focus on one or two topics, CHARMemes is multilingual and covers a broad set of topics. Finally, whereas several benchmarks frame the task as binary prediction, CHARMemes supports fine-grained harmfulness modeling with 11 categories, facilitating more detailed evaluation and error analysis.

5 Experiments and Results

Recent shared tasks on multimodal hate speech and meme understanding (Fersini et al., 2022; Dimitrov et al., 2024) show that both fine-tuned CNN/transformer architectures and lightweight classifiers trained on pretrained visual features are strong baselines. In particular, several works extract fixed representations from pretrained models such as CLIP (Radford et al., 2021) and train simple downstream classifiers (e.g., logistic regression or boosting), achieving competitive performance at substantially lower training cost (Zhang and Wang, 2022; Chen and Chou, 2022). Standard CNN backbones also remain effective for meme-related classification (Chikoti et al., 2024). Motivated by these findings, we evaluate representative baselines to assess whether CHARMemes supports meaningful learning and to study error patterns across modeling choices, rather than to optimize for state-of-the-art results.

5.1 Experimental Setup

Modeling conditions. We consider three complementary setups: (i) **Text-only:** We classify meme caption text with XLM-RoBERTa (Conneau et al., 2020), motivated by evidence from SemEval-2022 Task 5 that RoBERTa-style models are strong for meme text analysis (Fersini et al., 2022). (ii) **Visual-only (fine-tuned CNN):** We fine-tune EfficientNet-B2 (Tan and Le, 2019) as a strong unimodal image baseline. (iii) **Visual-only (frozen features):** We extract CLIP embeddings (Radford et al., 2021) as fixed image representations and train lightweight classifiers on top, using Logistic Regression and XGBoost (Zhang and Wang, 2022; Chen and Chou, 2022).

Training details. Trainable neural models (XLM-RoBERTa and EfficientNet-B2) are fine-tuned for 5 epochs with a learning rate of 2×10^{-5}. For the frozen-feature setup, CLIP is used only for feature extraction; Logistic Regression and XGBoost are trained as downstream classifiers (no end-to-end fine-tuning).

Evaluation protocol and metric. Given the class imbalance in CHARMemes, we report **Macro-F1** as the primary metric. We evaluate performance at three levels of granularity:

- **Binary classification:** Harmful vs. Safe.
- **Domain classification:** High-severity, Mid-severity, and Contextual harms vs. Safe (4 classes), as shown in Table 1.
- **Fine-grained classification:** The 11 categories defined in our taxonomy.

5.2 Harmful Meme Classification

Model Name	Fine-grained	Domain	Binary
Random	9.09	25.00	50.00
Majority Class	7.42	20.42	40.84
XLM-RoBERTa	<u>27.98</u>	<u>41.29</u>	58.82
EfficientNet	25.03	38.40	63.00
CLIP + LogReg	**25.53**	36.98	56.28
CLIP + XGBoost	24.95	**38.44**	**63.33**

Table 3: Macro F1-Scores for classifiers on different levels of granularity. Best results are **bold**, while best XLM-RoBERTa results are <u>underlined</u>, since it uses a slightly different training and testing sets.

Table 3 indicates that the fine-grained 11-class setting is challenging across all baselines, with Macro-F1 remaining below 30. XLM-RoBERTa attains the best Fine-grained (27.98) and Domain (41.29) scores, but its text-only formulation restricts evaluation to the captioned subset, excluding roughly 20% of CHARMemes. We therefore report it as a strong reference point, while prioritizing models that operate on the full dataset in the experiments that follow.

Among image-only approaches, CLIP-based visual features paired with lightweight classifiers are competitive: Logistic Regression yields the highest Fine-grained score (25.53), whereas XGBoost performs best on Domain (38.44) and Binary (63.33).

For consistency across levels of granularity, we derive the Domain (4-class) and Binary evaluations by mapping the 11-class predictions to their corresponding domain labels (Table 1) and to Harmful vs. Safe. Consequently, confusion among closely related Fine-grained categories does not necessarily imply failures at the Domain level. Our error analysis (Appendix B) supports this interpretation: misclassifications often occur within the same domain (e.g., between contextual harms), rather than

across domains. This helps explain why models such as CLIP+XGBoost can show modest Fine-grained performance while still making reasonably accurate safe/unsafe decisions.

5.3 Cross-Dataset Generalization

To further assess the utility of CHARMemes, we run cross-dataset transfer experiments in which a model trained on one dataset is evaluated on another. We use the best-performing binary classifier from Section 5.2 (CLIP+XGBoost) and select comparison datasets from Table 2. Since HarMeme (Pramanick et al., 2021a) contains two distinct topical subsets, we report results separately for Harm-P (US election-related) and Harm-C (COVID-related). For comparability, Table 4 reports four training configurations for each target dataset: **Base** (train/test on the target), CHARMemes **(Aligned)** (train on a size- and time-matched subset of CHARMemes), CHARMemes **(Full)** (train on the full CHARMemes training split), and *Reverse* (train on the target and test on a size-matched portion of CHARMemes). We focus on the binary setting because most benchmark datasets provide only harmful vs. safe labels (Table 2).

Table 4 reveals a consistent gap between in-domain and cross-domain performance. Training on CHARMemes transfers reasonably when using the full training split: CHARMemes (Full) matches or closely approaches Base on Harm-C (74.73 vs. 78.68) and MultiOFF (59.03 vs. 60.31), and remains competitive on Harm-P (54.81 vs. 56.00), while the drop is larger on the Hateful Memes Challenge (54.22 vs. 58.18). In contrast, CHARMemes (Aligned) typically underperforms both Base and CHARMemes (Full), suggesting that matching only time span and sample size is insufficient to counteract domain shift. A likely contributing factor is that the aligned subset is obtained via random sampling within the target time window, which does not guarantee comparability in topic mix, meme templates, or visual styles, and a more targeted selection strategy such as one suggested by Bates et al. (2025) could improve transfer. Finally, the **Reverse** direction is consistently harder: when training on the target datasets and testing on CHARMemes, the drop in results is most notable. The key factor for this could be a high amount of memes in Russian in CHARMemes, which makes generalization from English only way harder. Overall, our results suggest that CHARMemes can be considered as a useful training resource for creating detectors of harmful memes outside of Runet.

5.4 Temporal Generalization

Unlike many multimodal safety benchmarks that reflect short-lived events or are collected within a narrow time window, memes evolve continuously: templates appear and disappear, visual styles shift, slang and coded references drift, and changing cultural contexts reshape what "harmful" looks like in practice. This raises a key question: do models trained on past meme distributions remain reliable on future memes, or do they become outdated as the ecosystem changes? To study this, we evaluate *temporal generalization* by training on historical slices of the dataset and testing on different time periods.

We partition the dataset into five *consecutive* temporal splits—2007–2016, 2016–2017, 2017–2019, 2019–2020, and 2020–2025. Although these intervals are uneven in duration, they are defined to keep the number of examples and class distribution per split nearly constant (approximately 4.6K samples each), which helps isolate temporal distribution shift from changes in training-set size. To measure temporal generalization, we run a cross-period evaluation protocol: for each split S_i, we train a model using only 90% of examples from S_i and then evaluate it on every other split S_j with $j \neq i$, and the rest 10% of split S_i. This yields a full train–test matrix over time, allowing us to quantify how performance changes with increasing temporal distance and to compare how well different model families transfer from earlier to later meme distributions.

Our results in Figure 3 suggest that temporal generalization is challenging for fine-grained labels, whereas binary harmfulness detection is substantially more stable over time. In particular, CLIP-feature-based models generally maintain reasonable binary macro F-score when trained on a particular period and evaluated on the other ones, even when they struggle to preserve fine-grained distinctions. This matches the intuition that broad safe/unsafe cues generalize better than taxonomy-level labels, which are more sensitive to emerging formats and shifting cultural references. Results for other model families appear in Appendix K.

6 Conclusion and Future Work

We introduced CHARMemes a large-scale dataset for harmful meme detection in the Russian-oriented

Experiment Setup		Training Set Statistics				Binary Results		
Target dataset (test)	**Training data**	**Time Span**	**#Train / #Test**	**#Harm (train)**	**#Safe (train)**	**Accuracy**	**Macro-F1**	**ΔF1**
Hateful Memes Challenge	Hateful Memes (Base)	2020	8,500 / 1,000	3,089	5,481	61.30	**58.18**	0.00
	CHARMemes (Aligned)	2007–2025	8,500 / 1,000	3,089	5,481	55.00	54.13	-4.05
	Reverse (train Hateful Memes)	2020	8,500 / 1,000	3,089	5,481	52.60	37.94	-25.39
	CHARMemes (Full)	2007–2025	20,719 / 1,000	6,419	14,300	55.50	54.22	-3.96
Harm-C	Harm-C (Base)	2020–2021	3,013 / 354	1,064	1,949	79.66	**78.68**	0.00
	CHARMemes (Aligned)	2019–2022	3,013 / 354	1,061	1,952	66.38	58.41	-20.27
	Reverse (train HC)	2020–2021	3,013 / 354	1,064	1,949	58.19	46.41	-16.92
	CHARMemes (Full)	2007–2025	20,719 / 354	6,419	14,300	76.55	74.73	-3.95
Harm-P	Harm-P (Base)	2020–2021	2,938 / 355	1,488	1,450	56.06	**56.00**	0.00
	CHARMemes (Aligned)	2019–2022	2,938 / 355	1,475	1,463	53.52	53.04	-2.96
	Reverse (train HP)	2020–2021	2,938 / 355	1,488	1,450	53.80	50.99	-12.34
	CHARMemes (Full)	2007–2025	20,719 / 355	6,419	14,300	56.06	54.81	-1.19
MultiOFF	MultiOFF (Base)	2016	445 / 149	187	258	62.16	**60.31**	0.00
	CHARMemes (Aligned)	2016	445 / 149	183	262	59.46	49.40	-10.91
	Reverse (train MO)	2016	445 / 149	187	258	55.70	52.08	-11.25
	CHARMemes (Full)	2007–2025	20,719 / 149	6,419	14,300	59.06	59.03	-1.28

Table 4: Cross-domain **binary** classification results (Harmful vs. Safe) across four target datasets. For each block, the **Target dataset (test)** is fixed, and we vary the **training data**: (i) **Base** trains and tests on the target dataset; (ii) CHARMemes (Aligned) trains on a subset of CHARMemes matched to the target dataset in both time span and training-set size; (iii) *Reverse* trains on the target dataset and tests on CHARMemes; and (iv) CHARMemes (Full) trains on the full CHARMemes training set. **Time Span** denotes the collection period of the training source, and #Harm/#Safe report class counts in the training split. Results are reported as percentages. ΔF1 is computed relative to the Base Macro-F1 within each block; for *Reverse*, ΔF1 is relative to 63.33 (in-domain CLIP+XGBoost on CHARMemes; Table 3).

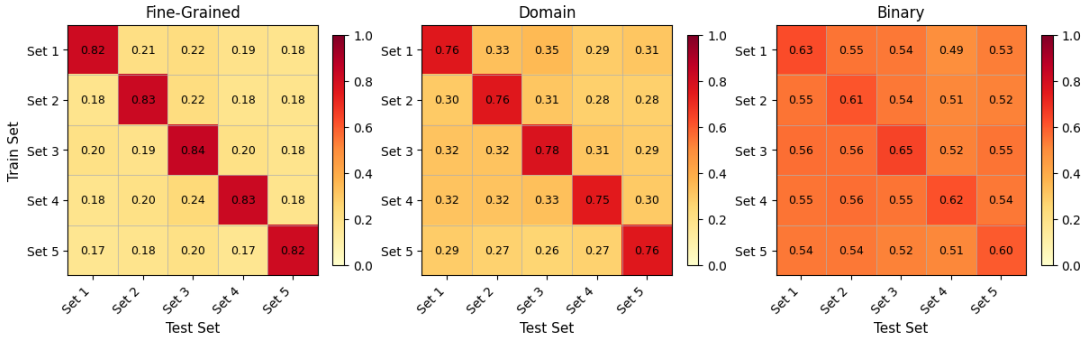


Figure 3: Temporal generalization of CLIP+XGBoost model.

online ecosystem (Runet). It aggregates in-the-wild memes from a template-centric public source and covers a broad historical span, enabling both cross-domain and temporal analyses of harmful meme content. To support fine-grained safety research, we proposed a harm taxonomy with 11 categories, additionally organized into three severity buckets (high-severity, mid-severity, and contextual/platform-policy harms), and designed an annotation workflow that combines strong automated labeling with targeted human review and extensive deduplication. Our experiments showed that fine-grained harmfulness recognition remains challenging, while binary harmful-vs-safe detection is substantially more robust, including under temporal and cross-dataset transfer. These results highlight both the progress and the remaining areas for improvement of current approaches when faced with cultural context, evolving meme formats, and shifting topical distributions.

We hope CHARMemes helps drive progress in harmful meme detection, supporting safer online spaces and more reliable, respectful communication across languages, communities, and evolving meme trends.

In future work, we plan to extend the dataset to include memes from platforms beyond Runet. We also intend to focus more on modeling. While the current work primarily introduces the dataset and highlights the limitations of existing approaches, our next steps will be to develop new methods for harmful meme detection that are more robust for fine-grained classification across different time periods and media formats.

Limitations

Despite the contributions of this work, several limitations should be acknowledged.

First, the primary focus of this study is on dataset creation rather than on developing state-of-the-art classification models. Consequently, model selection and training were not exhaustively explored. While we followed best practices reported in prior shared tasks, the chosen models may not fully reflect the performance potential achievable with more extensive experimentation and optimization.

Second, the notion of *harmfulness* in memes is inherently subjective. To mitigate annotator bias, each sample was labeled by multiple annotators, and both human annotators and large language models were employed. At the same time, our human annotators were active users of Runet communities, which may have influenced their judgments. As a result, individuals from different cultural backgrounds or age groups may perceive and label the same content differently.

Finally, as a resource paper, our primary goal is to provide a scalable, well-documented benchmark; in principle, such datasets benefit from comprehensive human validation. However, full re-annotation of the entire dataset would be prohibitively costly and would not materially improve the benchmark's utility given our quality controls. Since we observe consistently decent inter-annotator agreement and strong alignment between the annotators and the subset of GPT-5 labels we audited (Appendix E), we do not perform a full human evaluation over all samples.

Ethical Statement

While our dataset is designed to support research on meme understanding and the detection of harmful content in multimodal systems, we acknowledge the potential for misuse. The dataset contains real-world memes that may include hateful, harassing, or offensive language, as well as harmful stereotypes and sensitive cultural references directed at individuals and social groups. As a result, there is a risk that the dataset could be used to train or prompt models to generate or amplify harmful memes, targeted harassment, or disinformation. We explicitly discourage any such use and release the dataset for academic research purposes aimed at improving the safety, robustness, and accountability of vision–language and language models. We further note that meme interpretation is highly context-dependent: captions and visuals can be benign in one context but harmful in another, and some examples may be disturbing to annotators and readers. Also, given the sensitive topics covered in the dataset, annotators provided informed consent before annotation and had an option to withdraw at any time.

Finally, we used all datasets and web-sites in compliance with their usage rules and regulations, but since memes often derive from copyrighted material and may depict identifiable individuals, we encourage downstream users to respect source licensing and to avoid applications that enable harassment or content amplification. The dataset is intended for harm-aware analysis and evaluation, not for meme generation.

References

Florian Arendt, Sebastian Scherr, and Daniel Romer. 2019. Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults. *New Media & Society*, 21(11-12):2422–2442.

Luke Bates, Peter Ebert Christensen, Preslav Nakov, and Iryna Gurevych. 2025. A template is all you meme. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10443–10475, Albuquerque, New Mexico. Association for Computational Linguistics.

Antônio F. Carvalho, Igor Oliveira B. de Morais, and Thamyrys B. Souza. 2023. Profiting from cruelty: Digital content creators abuse animals worldwide to incur profit. *Biological Conservation*, 287:110321.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "subverting the jewtocracy": Online antisemitism detection using multimodal deep learning. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, page 148–157, New York, NY, USA. Association for Computing Machinery.

Lei Chen and Hou Wei Chou. 2022. RIT boston at SemEval-2022 task 5: Multimedia misogyny detection by using coherent visual and language features from CLIP model and data-centric AI principle. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 636–641, Seattle, United States. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Cheng Yu, and Jingjing Liu. 2019. Uniter: Universal

image-text representation learning. *arXiv preprint arXiv:1909.11740*.

Shreenaga Chikoti, Shrey Mehta, and Ashutosh Modi. 2024. IITK at SemEval-2024 task 4: Hierarchical embeddings for detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1779–1787, Mexico City, Mexico. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 2098–2110, New York, NY, USA. Association for Computing Machinery.

Jakob Johan Demant and Kristoffer Magnus Bjerre Aagesen. 2022. *An analysis of drug dealing via social media: Background paper commissioned by the EMCDDA*. EMCDDA.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026, Mexico City, Mexico. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. 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 6603–6617, Online. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4):1–30.

Hexiang Gu, Qifan Yu, Saihui Hou, Zhiqin Fang, Huijia Wu, and Zhaofeng He. 2025. Mememind: A large-scale multimodal dataset with chain-of-thought reasoning for harmful meme detection. *Preprint*, arXiv:2506.18919.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PLOS ONE*, 16(8):e0256762.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Competition report. In *Proceedings of Machine Learning Research (PMLR)*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*.

Etienne G Krug, James A Mercy, Linda L Dahlberg, and Anthony B Zwi. 2002. The world report on violence and health. *The lancet*, 360(9339):1083–1088.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of chinese harmful memes. In *Advances in Neural Information Processing Systems*, volume 37, pages 13302–13320. Curran Associates, Inc.

Clare McGlynn and Erika Rackley. 2017. Image-based sexual abuse. *Oxford Journal of Legal Studies*, 37(3):534–561.

Angela Nagle. 2017. *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right*. John Hunt Publishing.

Shraman Pramanick, Dimitar Dimitrov, Rumi Mandal, Shanu Krishnan, Md Shad Akhtar, and Preslav Nakov. 2021a. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hugo Rosa, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *Preprint*, arXiv:1910.02334.

Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A framework of severity for harmful content online. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Yukari Seko and Stephen P Lewis. 2018. The self—harmed, visualized, and reblogged: Remaking of self-injury narratives on tumblr. *New media & society*, 20(1):180–198.

Chhavi Sharma, Deepesh Bhageria, William Scott, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. Semeval-2020 task 8: Memotion analysis – the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020)*. Association for Computational Linguistics.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. *Preprint*, arXiv:2205.04274.

Shivam Sharma, Atharva Kulkarni, Tharun Suresh, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. Characterizing the entities in harmful memes: Who is the hero, the villain, the victim? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2149–2163, Dubrovnik, Croatia. Association for Computational Linguistics.

Limor Shifman. 2013. Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of Computer-Mediated Communication*, 18:n/a–n/a.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020a. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC)*. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John P. McCrae, and Paul Buitelaar. 2020b. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Surendrabikram Thapa, Aditya Shah, Farhan Jafri, Usman Naseem, and Imran Razzak. 2022. A multimodal dataset for hate speech detection on social media: Case-study of russia-Ukraine conflict. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 1–6, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267.

Marc Tuters and Sal Hagen. 2020. (((they))) rule: Memetic antagonism and nebulous othering on 4chan. *New media & society*, 22(12):2218–2237.

Emily A Vogels. 2021. *The state of online harassment*, volume 13. Pew Research Center Washington, DC.

Bradley E Wiggins. 2019. *The discursive power of memes in digital culture: Ideology, semiotics, and intertextuality*. Routledge.

Janis Wolak, David Finkelhor, Wendy Walsh, and Leah Treitman. 2018. Sextortion of minors: Characteristics and dynamics. *Journal of Adolescent Health*, 62(1):72–79.

Jing Zhang and Yujin Wang. 2022. SRCB at SemEval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 585–596, Seattle, United States. Association for Computational Linguistics.

Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2021. Racist or sexist meme? classifying memes beyond hateful. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 215–219, Online. Association for Computational Linguistics.

A Dataset Statistics

Figure 4 shows the class distribution in our dataset. Overall we have 23,025 samples. The overwhelming majority of safe samples is reasonable, since this reflects real-world platform distributions, where most content is non-harmful and only a smaller fraction falls into clearly harmful categories. However, the resulting class imbalance also makes the task more challenging: models can achieve deceptively strong overall accuracy by over-predicting the majority class, while still performing poorly on rare but high-impact harm categories. For this reason, in addition to reporting aggregate metrics, we emphasize class-aware evaluation (e.g., macro-averaged scores) and analyze per-class performance to better capture fine-grained detection behavior.

In terms of language distribution in the dataset, 42.77% of all image captions are in Russian. 24.22% are in English, 20.35% of samples do not contain any caption and the rest of the data is in other languages, including but not limited to: Ukrainian, Belarusian, Kazakh, etc. There are on average 7-13 words in caption. The TF-IDF analysis reveals strong lexical separation between classes, suggesting that captions contain category-specific cues rather than generic language. Offensive, Harassment, and Hate Speech are driven by strongly polarizing, targeted expressions, while Self-Harm is characterized by mental-health and self-harm indicators (e.g., "suicide", "depression", "grave"). Propaganda is dominated by political entities and election-related terminology (e.g., "elections", "Russia", "Navalny", "Putin", "Trump"), and Illegal Content is strongly linked to substance-related vocabulary.

The image set is dominated by relatively small, web-style meme resolutions, with a pronounced mode at 360×270 and a median size of 500×397 (approximately 0.22 MP). Aspect ratios are largely "standard" (median and 75th percentile at 1.33, i.e., close to 4:3), while a long tail of wide images (up to 9.10) suggests the presence of panoramas, multi-panel collages, or elongated screenshots. File formats are primarily JPEG (21,951 `.jpg` + 444 `.jpeg`, ∼79%), with a sizeable PNG minority (∼20%), consistent with a mix of photographic content and text-heavy graphics. There are notable outliers - very large files (up to ∼14.7 MB) and extremely high Laplacian variance—which likely correspond to high-resolution screenshots or detailed graphics.

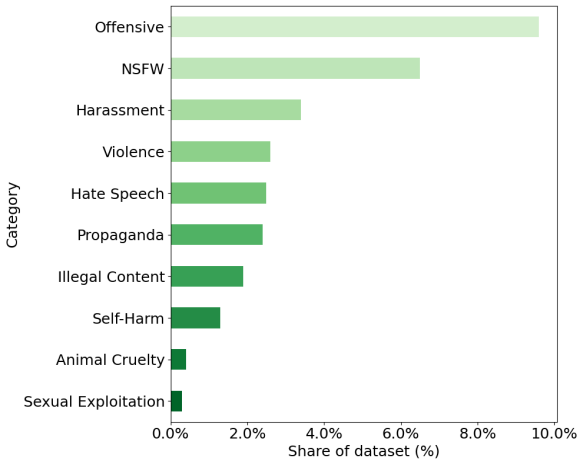


Figure 4: Category distribution in the dataset. The rest are Safe samples.

B Error Analysis

Confusion matrix on Figure 5 shows that misclassification often happens within the same harm domain. It explains why performance for 4 and 2 classes is better than for fine-grained 11-class classification.

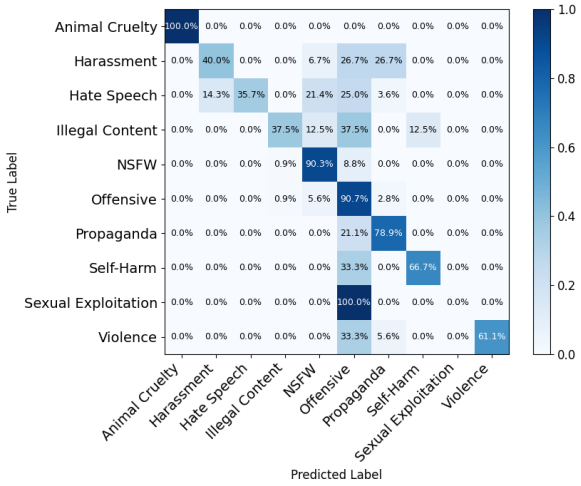


Figure 5: Confusion-Matrix for Harmful Meme Classification (only unsafe classes are used for clarity).

C Annotator's Guidelines

To ensure a high-quality and standardized format for instruction-output annotations, we provide clear guidelines throughout the annotation process. In this section we provide guidelines given to annotators before they started work. The text of the annotation shared with annotators is given below.

For convenience, annotators were provided with an annotation tool, see in Appendix G. They were also encouraged to use web-search and translation tools in cases when memes hard to understand or they were in a language other than Russian or English.

Meme Content Annotation Guidelines

Task. Annotators label Russian-language memes into content-moderation categories (harmful/offensive/policy-violating). When multiple harms appear, the final label reflects the **primary harmful content**.

Two modes. **Validation** measures agreement on a shared subset; **Labeling** assigns final labels to challenging cases where model suggestions disagree.

Severity hierarchy. If multiple categories apply, choose the **most severe**:

```
Sexual Exploitation > Violence >
       Self-Harm >
        Hate Speech
                    > Harassment >
  Animal Cruelty >
    Illegal Content
                    > Propaganda >
      Offensive >
      NSFW > Safe.
```

Label set.

- Safe: non-violating content.
- Hate Speech: attacks/dehumanization toward protected groups.
  - *Subcategories:* Racism & Xenophobia; Sexism; Homophobia & Transphobia; Religious Hate; Ableism; Targeted Hostility.
- Harassment: bullying/threats/doxing targeting identifiable individuals.
  - *Subcategory:* Cyberbullying & Doxing.
- Violence: depiction/promotion/glorification of physical harm or extremist violence.
  - *Subcategories:* Glorification of Violence; Terrorism & Extremism.
- Self-Harm: encouragement/glorification/instructions for self-injury or eating disorders.
  - *Subcategories:* Suicide & Self-Injury; Eating Disorders.
- Sexual Exploitation: non-consensual/coercive sexual content or any sexualized content involving minors.
  - *Subcategories:* CSAM; Trafficking; Sexual Coercion.
- Illegal Content: promotion/facilitation of illegal activities (e.g., drugs, scams).
  - *Subcategories:* Drugs & Gambling; Scams & Fraud.
- Propaganda: political misinformation intended to mislead (satire/opinion excluded).
  - *Subcategory:* Political Misinformation.
- Offensive: vulgarity, shock/gore, crude sexual humor, or cultural stereotyping without protected-group attacks.
  - *Subcategories:* Vulgarity & Profanity; Shock Content & Gore; Sexual Aggression/Humor; Cultural Stereotyping.
- NSFW: legal/consensual sexual or suggestive content inappropriate for general audiences.
  - *Subcategories:* Nudity/Suggestive; Adult Humor; Relationship Jokes.
- Animal Cruelty: depiction/promotion/glorification of intentional animal harm.

Decision rules.

1. Primary content: label the most severe harm (per hierarchy).
2. Text vs. image: text usually determines the meme message; *exception:* extreme violence or CSAM in the image dominates.

3. Surface-form labeling: satire/critique containing slurs or hateful statements is labeled as `Hate Speech` (models learn from visible cues).

4. Uncertainty: if unsure between two categories, choose the more severe; if the image is corrupted/blank, mark as invalid (Skip/Trash depending on mode).

5. Language: if text is not understood, use visual cues and translation if needed.

Summary Checklist. Before you begin, please make sure you:

- understand the difference between Validation and Labeling modes;
- have reviewed all category and subcategory definitions;
- know the severity hierarchy (Sexual Exploitation $\rightarrow \ldots \rightarrow$ Safe);
- understand when to use Skip vs. Trash;
- are prepared for exposure to potentially disturbing content and take breaks when needed;
- know how to report issues and provide progress updates.

D Annotation Prompt

In this section we provide prompt used for annotation with GPT-5.

Annotation Prompt

```
You are an expert meme classifier.
Your task is to analyze a meme image (including both
       its text and visuals) and classify it
      according to the taxonomy below.

You must return three things:
1. The text caption from the meme (OCR the text,
      exactly as shown).
2. One or more categories (from the taxonomy).
3. For each category, the most specific subcategory.

If no harmful content applies, return category "Safe
       / None".

---

### Taxonomy

 Hate Speech
   Content that expresses or promotes hatred,
         dehumanization, or discrimination against
         protected groups or individuals based on
         protected characteristics.

   **Racism & Xenophobia**
   - Attacks based on race, ethnicity, nationality,
         or immigration status
   - Racial slurs, stereotypes, or dehumanizing
         language
   - Content promoting racial superiority or
         segregation
   - Anti-immigrant sentiment or ethnic
         discrimination

   **Sexism**
   - Derogatory content about women or men based
         solely on gender
   - Gender-based stereotypes used to demean or
         belittle
   - Content promoting gender inequality or misogyny
         /misandry
   - Sexual objectification combined with
         dehumanization

   **Homophobia & Transphobia**
   - Attacks on LGBTQ+ individuals or communities
   - Content denying LGBTQ+ rights or humanity
   - Slurs or derogatory terms related to sexual
         orientation or gender identity
    "Conversion therapy" promotion or similar
          harmful content

   **Religious Hate**
   - Attacks on religious groups or beliefs
   - Anti-Semitism, Islamophobia, or hatred toward
         any faith
   - Content calling for violence against religious
         communities
   - Mockery intended to incite hatred (distinct
         from respectful critique)

   **Ableism**
   - Discrimination against people with disabilities
   - Mockery or dehumanization based on physical or
         mental conditions
   - Content suggesting disabled individuals are
         lesser or burdensome
   - Use of disability-related slurs as insults

   **Targeted Hostility**
   - Hate speech not fitting other subcategories
   - Attacks on age, appearance, or other
         characteristics
   - Coordinated harassment campaigns
   - Content explicitly calling for exclusion or
         harm

   **Important Distinction:**
   Hate speech requires intent to demean or attack a
          protected group. Crude jokes without
         malicious intent may fall under "Offensive"
         instead.

   Offensive
   Content that is crude, vulgar, or insensitive but
          lacks the targeted hatred characteristic of
          hate speech.

   ** Vulgarity & Profanity**
   - Heavy use of swear words or crude language
   - Sexually explicit jokes or innuendo (non-
         harmful)
   - Toilet humor or gross-out content
   - Content that would be inappropriate in
         professional settings

   ** Shock Content & Gore**
   - Disturbing imagery of death, injury, or
         violence
   - Graphic medical content or accidents
   - Content designed to shock or disgust viewers
   - "Edgy" humor relying on disturbing visuals

   ** Sexual Aggression/Humor**
   - Sexual jokes or innuendo that objectify without
          consent context
   - Crude sexual references or gestures
   - Content reducing individuals to sexual objects
   - Rape jokes or sexual violence humor (if not
         clearly promoting harm-otherwise classify as
          Hate Speech or another more severe category
          if applicable)

   ** Cultural Stereotyping**
   - Stereotypes about cultural practices, accents,
         or traditions
   - Mocking cultural dress, food, or customs
   - "Exotic" portrayals that otherize groups
   - Content perpetuating harmful national or
         regional stereotypes

 Safe / None
   - No hateful, offensive, or otherwise harmful
         content detected.

---

### Output Format (strict JSON)

{
  "caption": "<exact text from the meme>",
  "labels": [
    {
      "category": "Hate Speech",
      "subcategory": "2.1 Racism & Xenophobia"
    },
    {
      "category": "Offensive",
      "subcategory": "3.1 Vulgarity & Profanity"
    }
  ]
}
```

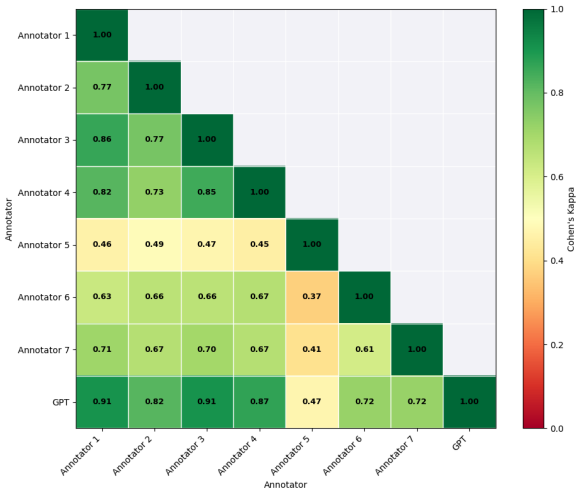


Figure 6: Inter-Annotator Agreement

E Inter-Annotator Agreement

As shown in Figure 6, the pairwise agreement between GPT-5 and the human annotators, measured with Cohen's κ, is generally high. One rater (Annotator 5) is a clear outlier with substantially lower agreement scores. A manual inspection of this annotator's labels suggested that many of them were incorrect, so we excluded this annotator from our final dataset. All reported results therefore rely on six annotators, whose Fleiss' κ is 0.721, indicating substantial agreement.

F Annotators' Demographics

We recruited six annotators. All were 21 years or older, had completed or were in the process of completing a higher-education degree, and had been subscribed to meme channels in Runet social networks since at least 2013. The group consisted of two women and four men.

G User Interface for Labeling

We provided annotators with a convenient labeling tool. Figure 7 shows the starting page of this tool, where annotators were allowed to pick the junk of data they are annotating, and Figure 8 shows labeling interface, where annotators were selecting the most appropriate class for the given label.

H Examples of memes from CHARMemes

In Table 5 we provide several samples of memes from the dataset.

[10] "Mukhosransk" is a derogatory Russian slang term for a remote, depressed provincial town (akin to " nowhere").

Content Warning: This section contains examples of memes that may be offensive, including hate speech and toxic content. These are included solely for research purposes. We by no means support the statements found in this section.

I Analysis of Labeling Disagreement

Table 6 provides examples of memes where the LLM (GPT-5) predictions across three independent runs differed. We also report the final human consensus label, selected for these memes.

J Covered Topics

Since our dataset covers a huge timeframe (from 2007 to 2025), it covers multiple information topics including, but not limited to: US elections, COVID, US sanctions, Russia-Ukraine conflict, Russian protests, constitutional reforms in Russia, climate change, the global financial crisis, the Crimea annexation, and the Belarus 2020 crisis. Across these years, memes react both to discrete headline events (e.g., elections, outbreaks, military escalations, major policy changes) and to longer-running sociopolitical processes (e.g., sanctions regimes, propaganda narratives, economic volatility, and shifts in public life). In addition, the dataset includes historically anchored themes that remain culturally salient in Russian online discourse, such as the USSR and Soviet nostalgia, as well as WWII / Great Patriotic War symbolism, commemorative culture, and their contemporary political and identity-related reinterpretations.

To capture this breadth without forcing overly narrow labels, we used a hybrid taxonomy: a compact set of recommended high-level topics plus an open-set mechanism for proposing new labels when the meme's theme did not fit the predefined list. Concretely, we used the following annotation prompt with GPT-5 assign topics to memes.

Topic Extraction Prompt

```
You are an expert meme analyst.
Task: extract meta-topic(s) from a meme image using
     BOTH text and visuals.

You have a recommended taxonomy below, but you MAY
     introduce new topics when needed.

Recommended taxonomy (ID -> name/definition):
{topics_block}

Return STRICT JSON only.

Output requirements:
- topics: up to 3 items, most relevant first
- Each topic item MUST be either:
  (A) a known topic: use "topic_id" from taxonomy
  (B) a new topic: use "new_topic" object (id/name/
```

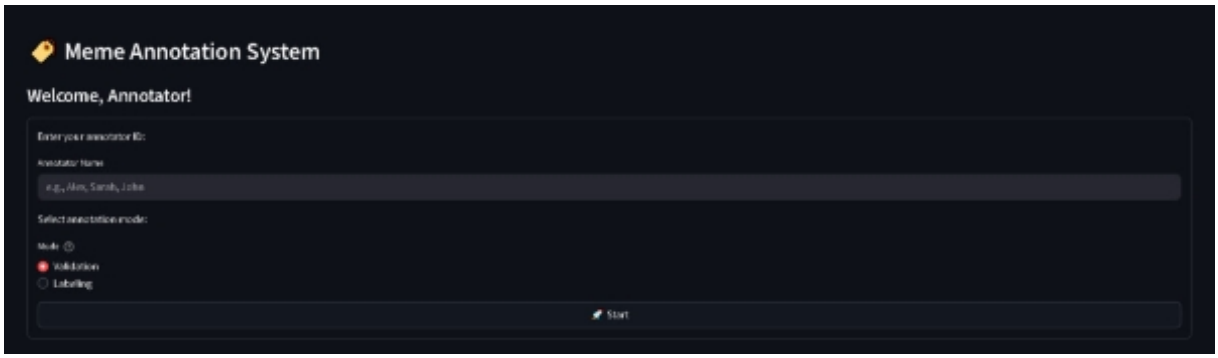


Figure 7: Welcome Page of the Labeling Interface

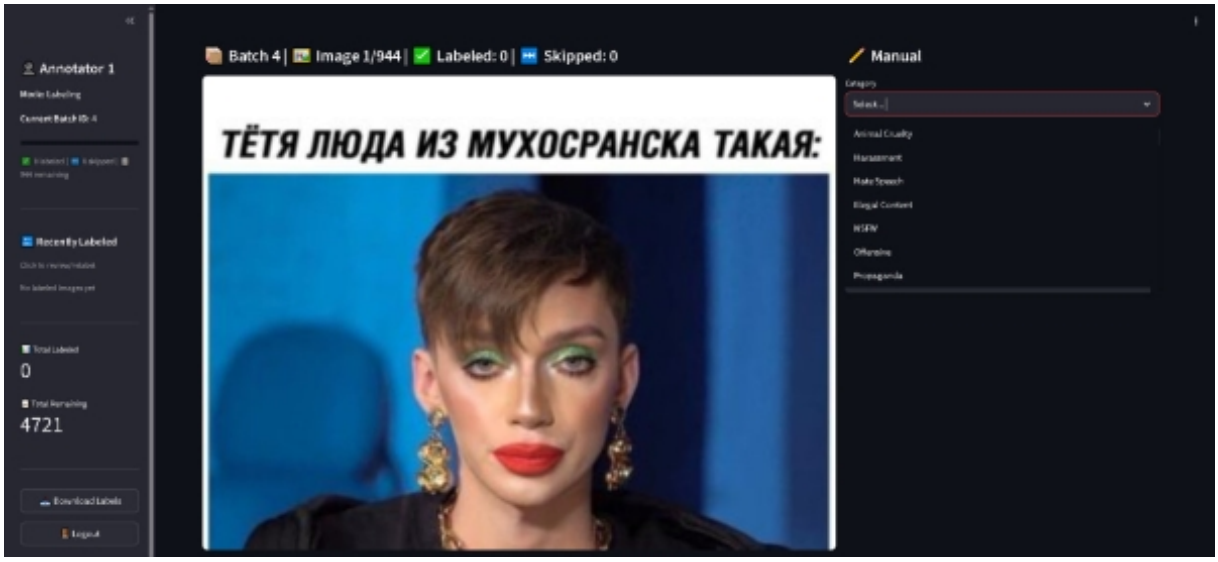


Figure 8: Main Labeling Interface for Annotators.
Translation: "Aunt Lyuda from Mukhosransk[10] be like..."

```
        definition)

STRICT OUTPUT JSON SCHEMA:
{
  "caption": "<exact text>",
  "topics": [
    {
      "kind": "known",
      "topic_id": "<one of taxonomy IDs>",
    },
    {
      "kind": "new",
      "new_topic": {
        "proposed_id": "short_snake_case",
        "name": "Human readable name",
        "definition": "What it covers (one sentence)
            ",
      },
    }
  ]
}

Hard constraints:
- Prefer kind="known" when a taxonomy topic fits
     reasonably.
- Only use kind="new" when it truly doesn't fit or a
      missing recurring theme is clear.
- topic_id must be exactly one of the taxonomy IDs
     listed above.
- proposed_id must be snake_case.
- Return JSON only (no markdown).
```

K Temporal Generalization Examples

Figure 9 shows a consistent temporal generalization pattern across models. Performance is highest when training and testing within the same historical split, and drops under cross-period transfer. EfficientNet exhibits the smallest apparent shift, largely because its overall performance is low across all splits. For the remaining models, temporal drift is clear for the 11- and 4-class settings, whereas binary harmfulness detection remains comparatively stable across time.

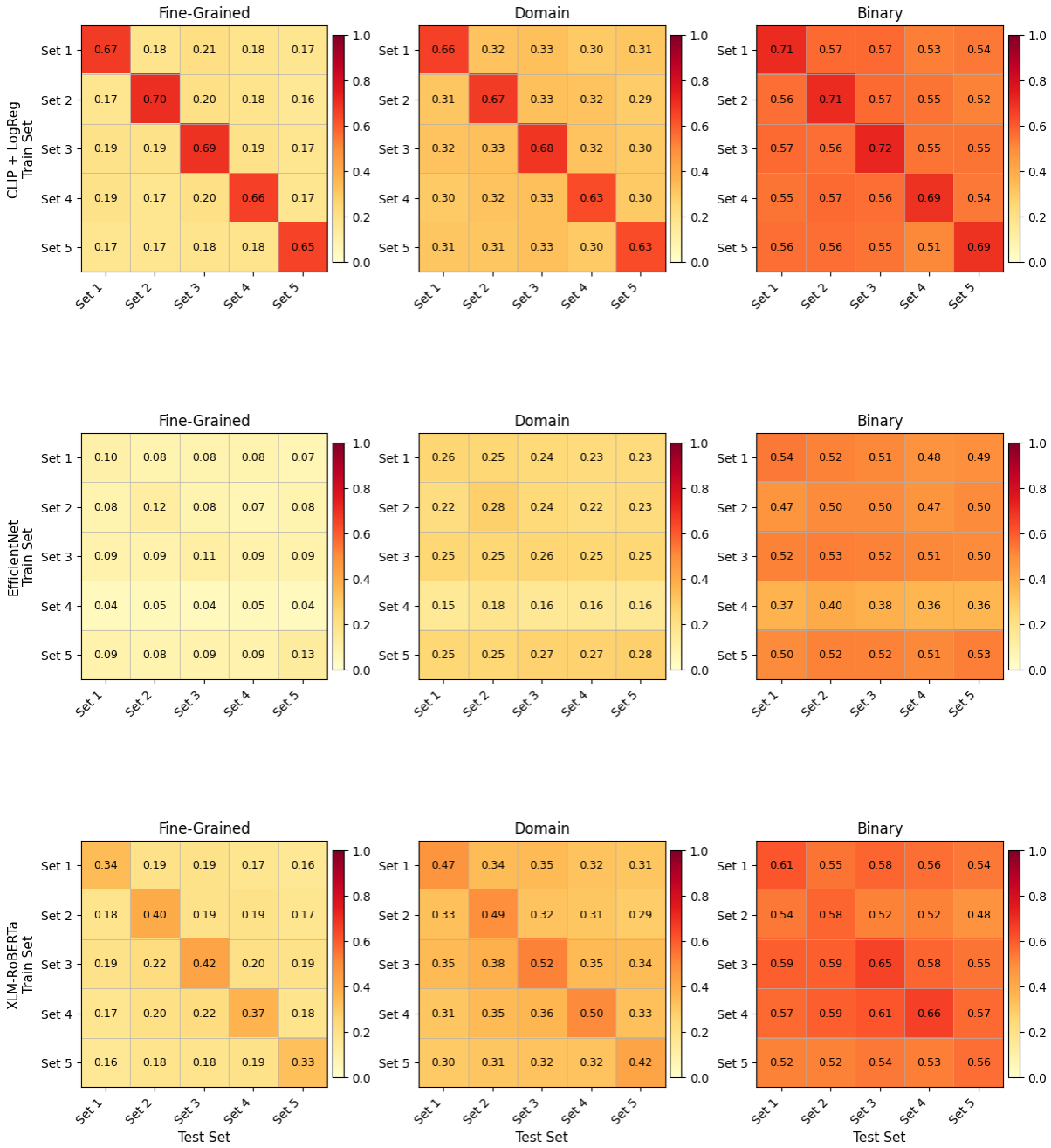


Figure 9: Temporal generalization of all models.

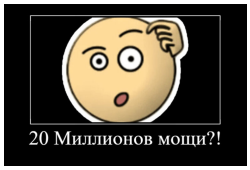
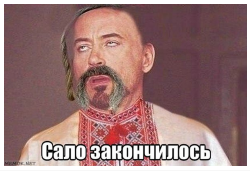
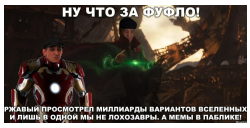
Meme Image	Original Caption	Translation	Class
	20 Миллионов мощи?!	20 Million powers?!	Safe
	*N/A*	*N/A*	Animal Cruelty
	Сало закончилось	We're out of salo	Offensive / Cultural-stereotyping
	Ну что за фуфло ржавый просмотрел миллиарды вариантов вселенных и лишь в одной мы не лохозавры. А мемы в паблике	What a load of cr*p. Rusty saw billions of timelines and there's only one where we aren't total losers. And the memes in this group	Offensive / Vulgarity & Profanity
	народ: хватит строить дачи, давай лучше дороги и больницы; неа	people: stop building mansions, give us roads and hospitals instead; nah	Propaganda / Political Misinformation

Table 5: Meme samples with corresponding labels and translation.

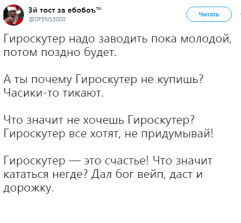
Meme Image	Original Caption	Translation	Labels by GPT (3 Runs)	Label by Human Annotators
	псс. парень за любовь выпить не хочешь?	psst. hey kid. wanna drink to love?	1. Illegal Content/Drugs & Gambling 2. Illegal Content/Drugs & Gambling 3. NSFW/Adult Humor	**Illegal Content/Drugs & Gambling**
	хочешь в питер?	wanna go to piter?	1. Self-Harm/Suicide & Self-Injury 2. Self-Harm/Suicide & Self-Injury 3. Violence/Glorification of Violence	**Self-Harm/Suicide & Self-Injury**
	Гироскутер надо заводить пока молодой, потом поздно будет. А ты почему Гироскутер не купишь? Часики-то тикают. Что значит не хочешь Гироскутер? Гироскутер все хотят, не придумывай! Гироскутер — это счастье! Что значит кататься негде? Дал бог вейп, даст и дорожку.	You need to get a hoverboard while you're still young, or it'll be too late. Why haven't you bought a hoverboard yet? The clock is ticking, you know. What do you mean you don't want a hoverboard? Everyone wants a hoverboard, stop making things up! A hoverboard is a blessing! What do you mean you have nowhere to ride? If god gave you a vape, he'll give you a pavement.	1. Safe 2. Safe 3. Offensive/Vulgarity & Profanity	**Safe**

Table 6: Examples of memes where automated annotation provided inconsistent labeling.