
Towards Fine-Grained Annotation of Directed Hate and Abusive Speech in Serbian

Miloš Košprdić¹[\[0000-0001-6902-3639\]](#), Isidora Dekić¹[\[0009-0005-6247-7669\]](#),
Tijana Radović¹[\[0000-0002-5759-0844\]](#)

¹University of Belgrade, Serbia,
{milos.kosprdic|isidora.r.dekic|tijana.n.radovic}@gmail.com

Scientific paper

DOI: 10.18485/judig.2026.2.ch23

Abstract

Hate-speech detection is challenging because abusive language, profanity, identity-based derogation, and group-directed hostility frequently overlap while remaining conceptually distinct. This distinction is particularly relevant in Serbian online discourse, where offensive and identity-related expressions may perform different pragmatic functions without necessarily constituting hate speech. We present a fine-grained annotation framework and accompanying guidelines for Serbian abusive and hate speech that explicitly model target structure, identity-based derogation, profanity function, quotation and reported abuse, and author stance. A 133,483-comment Serbian Reddit candidate corpus was constructed through phenomenon-oriented lexical retrieval and preprocessing, from which an intentionally enriched 500-comment pilot dataset was selected. The pilot was first processed through a structured LLM annotation pipeline and then annotated independently by multiple human annotators, enabling analysis of annotation reliability, construction of a human-consensus reference, and direct evaluation of the LLM first pass. The results support a clear distinction between abusive speech, identity-based derogation, and hate speech, while showing that the presence of profanity alone has little discriminatory value. Pragmatic and target-related judgments proved more difficult than those concerning explicit lexical phenomena, and the LLM showed a systematic tendency to apply the hate-speech label too broadly to abusive and identity-related content. The study therefore provides an operational annotation framework, accompanying guidelines, and a human-validated Serbian pilot resource, while establishing an empirical basis for refining LLM-assisted annotation and scaling resource construction to larger Serbian-language collections.

Keywords: Hate Speech, Abusive Language, Identity-Based Derogation, Fine-Grained Annotation, Serbian NLP, Human Annotation, LLM-Assisted Annotation

Warning: *This paper contains examples of potentially offensive content.*

1. Introduction

The rapid expansion of user-generated online content has increased the need for reliable detection of abusive speech and hate speech. A persistent difficulty, however, is that hate speech, abusive or offensive language, and profanity represent related but non-equivalent phenomena (Davidson et al., 2017; Vidgen et al., 2019).

This distinction is particularly relevant for Serbian, where profanity is a productive communicative resource and may perform a range of pragmatic functions, including emotional intensification, humor, irony, interpersonal aggression, and discourse marking (Živković Nikolić, 2024). This variability makes surface lexical offensiveness an unreliable basis for distinguishing abusive speech from hate speech.

Existing Serbian resources provide an important basis for abusive-language research but offer limited coverage of these distinctions. AbCoSER (Jokić et al., 2021), for example, differentiates several forms of abusive communication but does not explicitly model directedness, target structure, profanity function, quotation and reported speech, or author stance. These dimensions are particularly relevant to borderline cases such as identity terms used as interpersonal insults, reproduced hostile statements, and profanity-rich communication without group-directed hate.

We therefore propose a fine-grained annotation framework and accompanying operational guidelines for Serbian abusive speech and hate speech. The framework separates broad abusive communication from the narrower category of hate speech while additionally modeling directedness and target structure, identity-term presence and derogatory identity use, quotation and reported abuse, author stance, profanity and its pragmatic function, severity, and annotation uncertainty. Its central principle is that annotation should reflect the author’s communicative act rather than lexical offensiveness alone.

To develop and evaluate the framework, we first collected 171,337 raw Reddit comments through phenomenon-oriented lexical retrieval. After preprocessing, 133,483 comments formed the Serbian

Reddit candidate corpus, which is currently maintained as an internal research collection and serves as the candidate pool for subsequent annotation and resource expansion. From this corpus, we selected an intentionally enriched 500-comment pilot dataset. The pilot comments are first processed using a structured LLM annotation pipeline, producing what we refer to throughout the paper as *LLM first-pass annotations*, and are subsequently annotated independently by three human annotators per example. This design enables evaluation of both the reproducibility of the proposed annotation distinctions and the extent to which structured LLM first-pass annotations align with independent human judgments.

The contributions of this work are fourfold. First, we introduce a target-aware and pragmatically oriented annotation framework together with operational annotation guidelines for Serbian abusive speech and hate speech. Second, we construct the Serbian Reddit candidate corpus and derive from it a 500-comment pilot subset containing LLM first-pass and human annotations. Third, we empirically evaluate the applicability and reliability of the framework through 1,500 human annotation records, including both core and fine-grained dimensions. Fourth, we evaluate the LLM first-pass annotations against the resulting human-consensus reference and identify systematic model errors at the boundary between profanity, interpersonal abuse, identity-based derogation, and hate speech.

The remainder of the paper presents related work (Section 2), corpus and pilot-sample construction (Section 3), the annotation framework and guidelines (Section 4), and the annotation and validation procedure (Section 5) before reporting the quantitative (Section 6) and qualitative results (Section 6.5).

2. Related Work

A central challenge in hate-speech research is the operational distinction between hate speech and broader forms of harmful language. Previous work has repeatedly shown that profanity, offensive language, abusive speech, and hate speech should not be treated as interchangeable categories (Davidson et al., 2017; Waseem et al., 2017; Vidgen et al., 2019). Vulgar language may occur without group-directed hostility, while identity-directed hostility may be expressed without explicit slurs. Distinguishing these phenomena therefore requires information beyond lexical offensiveness, particularly about

the target of the utterance and the relevance of social identity (ElShrief et al., 2018).

Several influential resources have moved toward more structured annotation schemes. OLID (Zampieri et al., 2019) distinguishes offensive content from targeted offense and further specifies target type. FRENK (Ljubešić et al., 2019) demonstrated the importance of contextual and target-sensitive annotation in a multilingual setting, while HateXplain (Mathew et al., 2021) incorporated human rationales to support interpretable hate-speech modeling. These resources illustrate a broader shift from single binary labels toward annotations that represent how and toward whom harmful language is used.

For Serbian, AbCoSER (Jokić et al., 2021) represents the principal manually annotated abusive-language resource and distinguishes categories including profanity, derogatory language, and hate speech. It provides an important foundation for Serbian abusive-language research, but several distinctions relevant to pragmatically complex cases remain implicit. The present work builds on this foundation but adopts a different decomposition of the annotation problem. Rather than treating profanity, derogation, and hate speech primarily as subtypes of abusive content, it represents abusive speech and hate speech as core decisions alongside separate dimensions for target structure, identity-term presence and derogatory identity use, quotation and reported abuse, author stance, and the pragmatic function of profanity. This representation is designed in particular to capture cases in which identity-related or highly offensive language is present without group-directed hate.

These distinctions are particularly relevant in Serbian and related South Slavic online discourse because profanity and identity terminology may perform multiple pragmatic functions rather than consistently signaling group-directed hostility (Živković Nikolić, 2024). An identity-related slur, for example, may be used as an interpersonal insult without expressing hostility toward the corresponding group as such, while highly offensive language may be quoted or reported in order to criticize rather than endorse it.

Recent work has also examined LLMs as annotation assistants. LLM-generated annotations can reduce annotation effort and provide useful first-pass labels for subjective NLP tasks, but correspondence with human judgments varies across tasks and categories (Gilardi

et al., 2023). Such findings motivate human-in-the-loop approaches in which structured model annotations support resource construction while independent human annotation validates difficult conceptual boundaries. The present study evaluates this relationship directly by applying the same underlying conceptual framework in a structured LLM first pass and in independent human annotation.

3. Corpus and Pilot Sample Construction

3.1. Data Curation

The data curation process was designed to construct a Serbian-language Reddit candidate corpus enriched with potentially relevant examples of abusive language, identity-based derogation, and hate speech. Retrieval therefore intentionally favored phenomena relevant to the proposed distinctions.

Comments were collected through discussion threads retrieved from three Serbian-language Reddit communities: *r/serbia*, *r/AskSerbia*, and *r/serbiancringe*. Retrieval was guided by 19 manually constructed lexical groups covering profanity, general derogatory expressions, dehumanizing and intellectual-inferiority insults, gender- and sexuality-related insults, ethnicity and nationality terminology, religious references, ableist expressions, directedness indicators, and markers of implicit or exclusionary hostility. Lexical items were reduced to representative stems and combined into search queries; examples include *jeb-*, *idiot-*, *cigan-*, and *šiptar-*. When grouped searches yielded insufficient material, narrower queries or individual stems were used.

Retrieved discussions were subsequently re-matched locally against the original lexical stems to reduce search noise. The resulting raw collection contained 171,337 comments from 2,816 discussion threads and 23,480 unique accounts, covering the period from August 2012 to May 2026.

The collected comments underwent preprocessing that included Cyrillic-to-Latin transliteration and filtering of comments containing URLs or Markdown structures, emojis or common text emoticons, predominantly English content, or no word-level tokens. Comments above the 75th percentile of the length distribution were also excluded; this corresponded to a maximum length of 79 word-level tokens, with no minimum-length threshold. In total, 37,854 comments

were removed during preprocessing, leaving 133,483 for subsequent sampling.

The remaining comments were assigned heuristic candidate categories using lexical matches, co-occurrence patterns, and a priority hierarchy favoring identity- and hate-relevant phenomena. These categories represented potential protected-group hostility, interpersonal abuse, political or public-figure hostility, profanity without hostility, and non-hostile identity references. They were used exclusively to organize and control sampling and were never treated as annotation labels, human judgments, or reference categories.

3.2. Pilot Sample Construction

Early pilot iterations used stratified samples across the heuristic candidate categories to manually assess the linguistic diversity of the collected material and whether the proposed annotation dimensions captured the intended distinctions among harmful-discourse phenomena.

A predominantly proportional sampling strategy yielded relatively few comments involving identity-oriented hostility and difficult hate-speech boundary cases. For the final pilot, sampling was therefore restricted to the heuristic protected-group hostility candidate stratum. This produced an eligible pool of 746 comments, from which 500 comments were selected. This heuristic category was intentionally set broader than the operational hate-speech definition used in annotation and served only as a high-yield sampling criterion rather than as an annotation label.

Because both the original corpus retrieval and pilot selection were phenomenon-oriented, the resulting label distributions should be interpreted as properties of an intentionally enriched evaluation sample rather than as estimates of the prevalence of abusive speech or hate speech on Serbian Reddit.

4. Annotation Framework and Operational Guidelines

The proposed framework models potentially harmful discourse through complementary semantic, target-related, and pragmatic dimensions. Its central distinction is between broad abusive communication and the narrower category of hate speech. Additional dimensions are intended to explain why potentially offensive lexical material does or does not satisfy either category. Here, the term *framework* is

used in a methodological rather than ontological sense: it comprises the structured annotation scheme described in Section 4.1., the operational decision rules described in Section 4.2., and their application in the annotation and validation workflow described in Section 5.. It is not intended as a formal ontology.

4.1. Annotation Dimensions

Annotation is performed at the comment level, with each comment represented by one structured annotation record combining core labels with target-, identity-, quotation-, profanity-, and auxiliary dimensions. Table 1 summarizes these dimensions in grouped form.

Table 1: Annotation dimensions in the proposed framework.

Group	Dimensions	Purpose
Core labels	abusive, hate_speech	Broader abusive communication and narrower group-related hate speech.
Target	directed, target_present, target_type, protected_group	Presence, directedness, type, and social-identity category of the target.
Identity	identity_term_present, identity_term_used_derogatorily, derogated_identity	Presence, derogatory use, and category of identity-related terminology.
Quotation / stance	quoted_or_reported_abuse, author_endorses_quoted_abuse	Distinguishes reproduced abuse from the current author’s stance.
Profanity	profanity_present, profanity_function	Presence and pragmatic role of profanity.
Auxiliary	severity, uncertainty, evidence_span, rationale	Severity, annotation confidence, and supporting evidence or explanation.

In the annotation record, these dimensions are represented as comment-level attributes rather than separate span annotations for targets, identity-related expressions, or profanity. Thus, target pres-

ence, type, and directedness, identity-term presence and use, and profanity presence and function are recorded as properties of the complete comment. The only span-valued field is `evidence_span`, which identifies the relevant textual material supporting the annotation decision and may therefore include a target expression, identity-related terminology, profanity, or other decisive content.

4.2. Operational Guidelines for Hate Speech and Ambiguous Cases

Abusive speech is defined broadly as language that attacks, insults, humiliates, degrades, or threatens a target. Hate speech is defined more narrowly and is assigned only when two conditions are satisfied: (1) the target is a protected or socially identifiable group, or an individual targeted because of group membership; and (2) the current author’s communicative act expresses hostility, degradation, dehumanization, threat, exclusion, or harmful stereotyping toward that target.

Consequently, the presence of a slur or another identity-related term is not sufficient for a hate-speech label. Identity terminology used as a generalized interpersonal insult may be annotated as both derogatory and abusive while remaining non-hateful. This distinction is central to the framework because it separates identity-based interpersonal degradation from hostility directed toward a social group as such.

For example, if an individual is directly insulted through an identity-related expression, the annotation may assign `directed = true`, `target_present = true`, `target_type = individual`, `identity_term_present = true`, `identity_term_used_derogatorily = true`, and `abusive = true`. If the comment does not express hostility toward the corresponding social group or attack the individual on the basis of group membership, `hate_speech` remains false. The relevant target and derogatory expression are included in `evidence_span`.

Profanity is interpreted according to its pragmatic function rather than its lexical presence alone. The same lexical item may function as an attack, intensifier, discourse marker, humorous expression, quotation, or another pragmatic use. Profanity presence alone is therefore insufficient for determining abusive-speech or hate-speech status.

Annotation is based on the communicative act performed by the current author. For quoted or reported hostile content, the stance of the current author is distinguished from the stance expressed in the reproduced material. The presence of abusive or hateful language in quoted or reported material therefore does not by itself imply endorsement by the current author. Similar principles apply to irony, sarcasm, implicit hostility, political criticism, and cases with incompletely specified targets.

When the target, stance, or pragmatic interpretation cannot be determined confidently from the comment itself, uncertainty is recorded explicitly rather than resolved through unsupported external contextual assumptions. These operational principles form the core of the annotation guidelines and are intended to support consistent application of the framework to examples beyond the present pilot sample.

5. Annotation and Validation Pipeline

The pilot dataset was processed in two stages. First, all 500 comments received structured LLM first-pass annotations. Second, the same comments were independently annotated by human annotators in order to assess the reproducibility of the annotation dimensions, construct a human-consensus reference, and evaluate the LLM first-pass decisions.

5.1. LLM First-Pass Annotation

Each comment was processed independently without access to the surrounding Reddit thread. Annotations were generated using GPT-5.4 through the OpenAI Responses API. The zero-shot annotation prompt instantiated the framework definitions and operational principles described in Section 4., including the distinction between abusive and hate speech, target structure, identity-based derogation, quotation and author stance, profanity function, and the treatment of ambiguous cases. The model was instructed to annotate the current author’s communicative act rather than rely exclusively on surface lexical cues.

Outputs were constrained through structured generation and validated against a Pydantic-based schema corresponding to the annotation structure described in Section 4.1. Structurally invalid or incomplete outputs triggered automatic regeneration. Each retained

annotation therefore conformed to the predefined structure and contained all applicable annotation dimensions together with an evidence span, short rationale, severity and uncertainty estimates, and a `needs_human_review` indicator.

The review indicator was intended as an auxiliary prioritization signal for potentially difficult cases rather than as a substitute for human evaluation. The resulting model outputs represent the model’s initial application of the framework and are not treated as reference labels.

5.2. Independent Human Annotation

All 500 comments were subsequently annotated independently by three human annotators, yielding 1,500 human annotation records. Ten native Serbian speakers participated in the annotation study, with each annotator receiving 150 out of the 500 comments. Assignment was balanced so that every comment was annotated exactly three times while overlap between individual annotator pairs remained approximately uniform.

Human annotators followed the operational guidelines described in Section 4.2. but were not shown the LLM labels, rationales, uncertainty estimates, or human-review recommendations. As in the LLM first pass, comments were evaluated independently without access to the surrounding Reddit thread.

Annotations were collected in Google Sheets using a structured form implementing the framework dimensions applicable to human annotation. The form contained the categorical annotation fields together with `evidence_span`, rationale, severity, and uncertainty; the LLM-specific `needs_human_review` indicator was not included. Initially incomplete required decisions were returned to the corresponding annotators for completion. The finalized human dataset contains all 1,500 records with no unresolved required annotation decisions.

5.3. Agreement and Human-Consensus Reference

Inter-annotator agreement was evaluated separately for each annotation dimension. Because the task combines binary, nominal, ordinal, and hierarchically conditional variables, and because several labels exhibit strong prevalence imbalance, agreement was summarized using complementary measures rather than a single coefficient.

For the central dimensions, we report raw within-item pairwise agreement, the proportion of examples receiving unanimous three-of-three agreement, Krippendorff’s alpha, and Gwet’s AC1. The latter provides a complementary chance-corrected measure that is less sensitive to strong prevalence imbalance. Severity was evaluated using ordinal Krippendorff’s alpha. Conditional dimensions such as profanity function, derogated identity, and protected-group type were additionally evaluated only on annotations for which the corresponding parent condition was applicable.

For binary dimensions with three applicable judgments, a human-consensus label was obtained through majority vote. The central binary dimensions therefore received a determinate unanimous three-of-three or two-of-three label for every example. For multiclass dimensions, cases in which all three annotators selected different categories were retained as disagreement rather than artificially converted into a reference category.

Before fine-grained model–human comparison, semantically equivalent differences in category naming were normalized where a direct mapping existed, while categories without an equivalent counterpart were kept distinct rather than silently merged.

The LLM first-pass annotations were then evaluated against the human-consensus reference. Precision, recall, and F1 were used for the central binary dimensions, while multiclass analyses relied on accuracy and macro-F1 where appropriate. Performance was also examined separately for unanimous and two-of-three human decisions in order to distinguish model disagreement with comparatively stable human judgments from disagreement on more ambiguous examples.

6. Results

The results are organized according to the source and level of evidence produced by the annotation workflow. We first examine the distributions obtained from the LLM first-pass annotations independently of the subsequent human judgments. We then analyze the human annotations, focusing on the reproducibility of both the central and fine-grained dimensions, and derive a majority-based human-consensus reference for the pilot dataset. Only after the model and human annotations have been characterized independently do we compare the LLM first pass directly with human consensus and examine the resulting error patterns. Finally, representative examples

are used to illustrate the pragmatic and conceptual distinctions underlying the quantitative findings.

6.1. Initial LLM First-Pass Annotation Results

The structured LLM pipeline produced complete first-pass annotations for all 500 pilot comments. Table 2 summarizes selected dimensions central to the proposed annotation framework.

Table 2: Distribution of selected dimensions in the LLM first-pass annotations.

Dimension	Positive examples	Percent
Abusive speech	389	77.8%
Hate speech	142	28.4%
Profanity presence	461	92.2%
Identity-term presence	428	85.6%
Derogatory identity use	218	43.6%
Directedness	417	83.4%
Target presence	443	88.6%
Quoted/reported abuse	53	10.6%
Needs human review	213	42.6%

The first-pass annotations showed a pronounced asymmetry between abusive speech and hate speech. Of the 142 comments labeled as hate speech, 141 (99.3%) were also classified as abusive, whereas only 141 of the 389 abusive examples (36.2%) additionally received the hate-speech label.

A compact association analysis further characterized the internal structure of the first-pass annotations. Abusive-speech and hate-speech labels were moderately associated ($\chi^2(1) = 51.33$, $p < .001$, Cramér’s $V = .32$). By contrast, profanity presence showed only a weak and non-significant association with the hate-speech label ($\chi^2(1) = 1.75$, $p = .186$, $V = .06$), despite profanity occurring in most comments. The strongest relationship involved derogatory identity use, which was strongly associated with model-positive hate speech ($\chi^2(1) = 234.61$, $p < .001$, $V = .69$).

While profanity was nearly ubiquitous in the LLM annotations, it was assigned different pragmatic functions. The model classified profanity as an *attack* in 285 comments (57.0%), as an *intensifier* in 84 (16.8%), as a *discourse marker* in 36 (7.2%), and as *quotation* in

32 (6.4%); humorous and other uses were less frequent. The attack function was particularly dominant, occurring in 112 out of 142 hate-speech labeled cases (78.9%).

Target-related labels showed a consistent pattern among model-positive hate-speech examples. All 142 model-positive hate-speech examples were classified as directed and as containing an identifiable target; 125 received the target type *protected_group*. Explicit author endorsement of reproduced abusive material was rare. The model additionally marked 213 comments (42.6%) as requiring human review, providing a sizeable first-pass subset of cases identified as potentially difficult by the annotation pipeline.

6.2. Human Inter-Annotator Agreement

In assessing the reproducibility of the annotation dimensions, complementary agreement measures were used. Table 3 summarizes the four dimensions most directly related to the central research questions.

Table 3: Human inter-annotator agreement for the central annotation dimensions.

Dimension	Raw	3/3	α	AC1
Hate speech	.815	72.2%	.311	.747
Abusive speech	.724	58.6%	.418	.475
Profanity presence	.905	85.8%	.390	.888
Profanity function	.404	22.4%	.240	.329

Agreement differed considerably across dimensions. Profanity presence was the most reproducible central phenomenon, with .905 raw pairwise agreement and unanimous agreement on 85.8% of the examples. Hate speech also showed comparatively high observed consistency: 72.2% of comments received the same decision from all three annotators, while the remaining 27.8% received a two-of-three majority decision. Abusive speech proved less sharply delimited, with unanimous agreement on 58.6% of the sample.

For hate speech, the comparatively low Krippendorff’s alpha should be interpreted in the context of strong class imbalance: approximately 16% of individual hate-speech ratings were positive. Because expected agreement is correspondingly high, alpha was considered jointly with raw agreement, item-level unanimity, and Gwet’s

AC1 rather than as a standalone measure of reliability. These complementary measures indicate considerably greater practical consistency than the alpha coefficient alone would suggest.

A particularly strong contrast appeared between profanity presence and profanity function. Although annotators generally agreed on whether profanity occurred, assigning its pragmatic role was much more difficult. Only 22.4% of examples received unanimous function labels, 54.0% received a two-of-three majority, and 23.6% resulted in three different judgments. Restricting the analysis to annotations in which profanity was considered present did not eliminate the difficulty (raw agreement = .422). Thus, recognizing profanity as a lexical phenomenon was substantially more reproducible than interpreting its communicative function.

The fine-grained dimensions revealed a hierarchical pattern. Identity-term presence showed .811 raw agreement and $AC1 = .721$. Among annotations in which an identity term was considered present, raw agreement on whether it was used derogatorily was .723. Once derogatory identity use had itself been established, agreement on the corresponding identity category became substantially stronger ($\alpha = .609$; $AC1 = .685$).

For target-related annotation, raw agreement on whether an identifiable target was present was .644, while target-type agreement among target-present annotations was .546. Once the target was classified as a protected or socially identifiable group, however, agreement on the corresponding group category increased to .802 raw agreement ($\alpha = .728$; $AC1 = .759$).

Quoted or reported abuse was comparatively reproducible (raw agreement = .861). Among annotations in which reproduced abuse was considered present, observed agreement on whether the author endorsed it was .852. Positive endorsement itself was very rare, making prevalence-sensitive chance-corrected coefficients less informative for this field.

Severity showed moderate agreement when its ordinal structure was taken into account (ordinal Krippendorff's $\alpha = .438$). By contrast, annotation uncertainty showed almost no chance-corrected agreement ($\alpha = .029$), suggesting that uncertainty is better understood as an annotator-specific confidence signal than as a stable semantic property of the comment.

6.3. Human-Consensus Annotation Results

Majority aggregation of the three independent annotations produced a human-consensus reference for the central binary dimensions of all 500 examples. Table 4 summarizes selected dimensions of the resulting consensus annotations.

Table 4: Distribution of selected dimensions in the human-consensus annotations.

Dimension	Positive examples	Percent
Abusive speech	309	61.8%
Hate speech	67	13.4%
Profanity presence	474	94.8%
Identity-term presence	412	82.4%
Derogatory identity use	170	34.0%
Quoted/reported abuse	61	12.2%

The relationship between abusive speech and hate speech was strongly asymmetric. Of the 67 hate-speech examples, 64 (95.5%) were also classified as abusive, whereas only 64 of the 309 abusive examples (20.7%) were classified as hate speech.

Profanity presence offered little separation between hateful and non-hateful content, occurring in 61 of 67 hate-speech examples (91.0%) and in 413 of 433 non-hate-speech examples (95.4%). Its pragmatic function was more differentiated. Among profanity-present examples for which a human majority function label was available, *attack* was the most frequent category, followed by *intensifier*, *discourse marker*, *quotation*, and *humor*. Attack was dominant among hateful uses, whereas non-abusive comments more often contained intensifying, discourse-level, quoted, or humorous profanity.

Among the 67 human-consensus hate-speech examples, 55 (82.1%) contained derogatory identity use. However, another 115 comments containing derogatory identity use were classified as non-hateful. Target orientation was highly characteristic of human-consensus hate speech. All 67 human-consensus hate-speech examples contained an identifiable target, and 66 of 67 were classified as directed.

Quoted or reported abuse was identified in 61 comments overall, whereas explicit author endorsement of reproduced abusive content occurred only rarely.

6.4. LLM First-Pass Annotations versus Human Consensus

The degree of correspondence between the LLM first-pass annotations and human consensus varied across the central annotation dimensions. Table 5 reports precision, recall, and F1 for the three central binary dimensions.

Table 5: LLM first-pass annotations compared with human consensus.

Dimension	Human +	LLM +	P	R	F1
Hate speech	67	142	.415	.881	.565
Abusive speech	309	389	.746	.939	.831
Profanity presence	474	461	.985	.958	.971

The strongest correspondence was observed for profanity presence (F1 = 0.971). Abusive-speech annotation was also comparatively strong (F1 = .831), although the LLM assigned the positive label more often than the human consensus. The largest distributional difference, however, concerned hate speech: the LLM labeled 142 comments as hate speech compared with 67 under human consensus. This was reflected in high recall (.881) but substantially lower precision (.415).

The hate-speech error distribution further clarifies this pattern. The model produced 59 true positives, 350 true negatives, 83 false positives, and eight false negatives, indicating that the discrepancy was driven primarily by overclassification rather than by failure to identify human-positive cases. All eight false negatives occurred among comments whose human hate-speech decision was based on a two-of-three majority; the model did not miss any unanimously positive human hate-speech example. Conversely, 44 of the 83 false positives occurred despite unanimous rejection of the hate-speech label by all three human annotators. Model omissions therefore clustered in cases with human disagreement, whereas overclassification also occurred in many unanimously human-negative cases.

The profanity-function comparison showed a larger gap between lexical detection and pragmatic interpretation. Among 356 relevant examples with a human majority function label, LLM accuracy was .587 and macro-F1 was .429. The model disproportionately favored *attack*, assigning the function to 214 examples compared with 146 human-reference cases, while under-recognizing categories such as

intensifier and *discourse marker*. Performance was also higher on examples with unanimous human function labels than on two-of-three cases.

At the same time, the model performed considerably better on some lower-level identity classifications. On applicable cases, identification of the specific derogated identity reached .888 accuracy and .796 macro-F1, while protected-group classification reached .875 accuracy and .844 macro-F1.

6.5. Qualitative Analysis of Annotation Patterns and Errors

The quantitative results were further examined through qualitative analysis of the annotated comments. Table 6 presents four cases selected to illustrate both a clear instance of group-directed hate speech and several boundary phenomena central to the framework: identity-based interpersonal abuse, non-hostile pragmatic profanity, and reported identity-based abuse. Usernames and other identifying metadata are omitted, and the English glosses are provided as approximate translations.

The contrast between EX_0281 and EX_0018 illustrates the central boundary in the hate-speech decision. In EX_0281, hostility is directed toward an explicitly identifiable social group and forms part of the current author’s own communicative act. In EX_0018, by contrast, an identity-related expression functions as part of an interpersonal insult rather than as hostility toward the referenced group. The distinction therefore depends not on the mere presence of offensive or identity-related language, but on the target and role of that language in the author’s communicative act.

This contrast reflects the most consequential pattern in the LLM errors. Among the 83 false-positive hate-speech predictions, 82 contained an identity term according to human consensus, 60 involved derogatory identity use, 64 were classified as abusive, and 79 contained profanity. The model was therefore not responding merely to unrelated lexical noise: most false positives contained annotation-relevant signals such as identity terms, derogatory identity use, abusive language, or profanity. The main difficulty arose at the interpretative step of determining whether these signals reflected identity-related interpersonal degradation or hostility satisfying the narrower group-directed criteria for hate speech.

Table 6: Representative examples illustrating central annotation distinctions.

ID	Example	Annotations
EX_0281	<i>“Ovo nisu srbi vec hrvati iz Bosne, seme im ustasko jebem...” (“These are not Serbs but Croats from Bosnia; fuck their Ustaša seed...”)</i>	Human: hate speech TRUE (3/3), abusive TRUE (3/3); LLM: hate speech TRUE.
EX_0018	<i>“jebo vas taj cigan da vas jebe više smorili ste” (approximately: “fuck you and that [ethnic slur], you have become tiresome”)</i>	Human: hate speech FALSE (3/3), abusive TRUE, derogatory identity use TRUE (3/3); LLM: hate speech TRUE.
EX_0246	<i>“U jebote, that’s a first. Iznenadio bi se ... bukvalno imaš crkve pred kojima leti LGBT zastava...” (“Oh fuck, that’s a first. You would be surprised ... there are literally churches with LGBT flags outside...”)</i>	Human: hate speech FALSE (3/3), abusive FALSE (2/3), profanity function = discourse marker (2/3); LLM: hate speech FALSE, function = intensifier.
EX_0403	<i>“Bukvalno su me od osnovne škole deca zajebavala da sam šiptar ...” (“Children had mocked me since primary school by calling me [ethnic slur] ...”)</i>	Human: quoted/reported abuse TRUE, endorsement FALSE, hate speech FALSE (3/3); LLM: hate speech FALSE.

EX_0246 illustrates a different form of pragmatic ambiguity. The vulgar expression functions as a discourse-level reaction rather than as an attack on the identity group mentioned later in the comment. The disagreement between the human and LLM profanity-function labels in this example illustrates the difficulty of assigning pragmatic function even when profanity itself is readily identifiable.

Quotation and reported speech introduce a further source of difficulty. In EX_0403, an ethnic slur appears because the author reports abuse that had previously been directed toward them. The example therefore demonstrates why the occurrence of highly offen-

sive or identity-related lexical material cannot by itself determine the hate-speech annotation: it must also be established whether that material forms part of the current author’s own communicative act or is being reproduced in reported speech.

7. Discussion

The findings have implications at three complementary levels. First, they provide empirical evidence about the applicability and reliability of the proposed annotation framework and accompanying guidelines for Serbian harmful-language analysis. Second, they clarify the relationships among profanity, abusive speech, identity-based derogation, and hate speech and identify the pragmatic decisions that make these categories difficult to distinguish. Third, the comparison between model and human annotations provides an empirical basis for determining how structured LLM-assisted annotation can contribute to larger-scale Serbian resource construction without treating model outputs as authoritative labels.

7.1. Reinterpreting the LLM First Pass through Human Validation

The initial LLM first pass proved useful as an exploratory application of the annotation framework, but human validation substantially changed how its output should be interpreted. Although the model captured all unanimously positive human hate-speech cases, it operated with a substantially broader positive decision boundary and produced many false positives, including cases unanimously rejected by all three human annotators. This discrepancy remains meaningful even though neither set of label proportions represents prevalence in Serbian Reddit discourse, since the pilot corpus was intentionally enriched for relevant phenomena. The LLM first pass is therefore better interpreted as sensitive candidate detection than as autonomous reference annotation: structured model output may support screening and prioritization, but requires calibration and independent human validation.

7.2. Distinguishing Abuse, Profanity, and Identity-Based Derogation from Hate Speech

The human-validated results provide empirical support for the conceptual structure underlying the proposed annotation guidelines. Hate speech, abusive speech, profanity, and identity-based derogation overlap extensively but capture different properties of harmful communication.

Hate speech is substantially narrower than abusive speech. Most human-consensus hate-speech examples are simultaneously abusive, while the reverse relation holds for only a minority of abusive comments. Broader abuse therefore includes interpersonal insults, vulgar aggression, and hostile interaction that need not invoke group membership as the basis of hostility.

Profanity is broader still. Its near-ubiquity in the enriched sample and comparable occurrence in hateful and non-hateful comments demonstrate why lexical vulgarity alone provides little evidence for hate speech. What is more informative is the role profanity performs within an utterance, where the same lexical material may function as an attack, emotional intensifier, discourse marker, humorous device, quotation, or stylistic emphasis.

Identity-based derogation occupies a particularly informative intermediate position. It is frequent among hate-speech examples but also occurs in many comments classified as non-hateful. Identity terminology and slurs can function as interpersonal insults, generalized pejoratives, or components of hostile exchanges without necessarily expressing hostility toward the corresponding social group as such. Explicitly representing this distinction is therefore important both linguistically and computationally.

Across these distinctions, the principal difficulty lies not in identifying offensive or identity-related material as such, but in interpreting it in relation to its target, the relevance of group membership, its pragmatic function, and the current author’s stance. This in turn explains why the dimensions are represented separately in the framework rather than treated as interchangeable indicators of hate speech.

The annotation guidelines constitute an important part of this contribution. By operationalizing the conditions under which offensive, identity-related, quoted, or profane material should and should not be interpreted as hate speech, they provide a reproducible decision framework that can be applied beyond the present sample. At

the same time, the human agreement results identify dimensions – particularly target inference and pragmatic-function classification – for which additional examples and calibration could further improve annotation consistency.

7.3. Human Disagreement and Pragmatic Complexity

The human annotation study demonstrates that disagreement itself carries meaningful information about the task. Relatively observable phenomena such as profanity presence were annotated highly consistently, whereas dimensions requiring pragmatic interpretation showed greater variability.

The hierarchical analyses suggest that many difficulties arise at an applicability or “gate” stage. Annotators may disagree on whether a target or derogatory identity use is present while agreeing more strongly on the corresponding subtype once the phenomenon has been established. This observation has implications both for future annotation-guideline refinement and for computational systems that might benefit from explicitly hierarchical decision procedures.

Human disagreement should also be retained when automated systems are evaluated. A model decision that conflicts with all three human annotators represents a different type of error from one that differs from a two-of-three majority. The hate-speech results make this distinction particularly clear: all model false negatives occurred among majority vote cases, whereas many false positives contradicted unanimous human-negative judgments.

Finally, the very low agreement on annotation uncertainty suggests that annotators apply different personal standards when assessing confidence. Uncertainty may therefore remain useful as a metacognitive or review-prioritization signal, but it should not be treated as an objective semantic annotation dimension equivalent to abusive speech or hate speech.

7.4. Implications for LLM-Assisted Resource Construction

For larger-scale resource construction, the human-validated pilot provides an empirical basis for identifying which parts of the first-pass pipeline require recalibration before the workflow is applied to a substantially larger candidate pool.

The error profile also indicates where such recalibration should be concentrated. The model performed comparatively well in iden-

tifying specific derogated identities and protected groups, suggesting that it can often recognize which identity or group is involved. Its greater difficulty lies at the higher-level decision of whether offensive or identity-related hostility satisfies the narrower criteria for hate speech. Recalibration should therefore focus particularly on the hate-speech decision boundary and on the treatment of identity-based derogation, profanity function, and author stance.

The auxiliary `needs_human_review` signal also showed some potential for review prioritization: flagged examples tended to produce more human disagreement than unflagged examples. Such signals may therefore help allocate human annotation effort, although they cannot substitute for independent judgments.

Importantly, the present workflow also creates a pathway from the validated pilot to larger-scale Serbian resource construction. The 500-comment pilot was selected from a substantially larger collected Reddit corpus that remains available as a candidate pool for further annotation.

A revised pipeline could subsequently be applied to the larger collected corpus, while model-based review signals and stratified human validation are used to prioritize difficult cases and monitor annotation quality. In this sense, the study produces not only a validated pilot resource but also an empirically informed procedure for developing a substantially larger Serbian harmful-language dataset.

8. Conclusion and Future Work

This study introduced and empirically evaluated a fine-grained annotation framework and accompanying operational guidelines for abusive speech and hate speech in Serbian online discourse. A larger Serbian Reddit corpus was constructed as a candidate pool for harmful-language research, from which an intentionally enriched 500-comment pilot dataset was selected. Each pilot example received a structured LLM first-pass annotation and three independent human annotations, yielding 1,500 human annotation records and enabling the construction of a human-consensus reference while preserving information about annotator disagreement. The resulting pilot therefore combines a detailed annotation framework and guidelines with model-generated first-pass labels, independent human judgments, and empirical evidence about annotation reliability.

Overall, the human-validated results support the proposed separation of abusive speech, hate speech, profanity, and identity-based derogation, while identifying pragmatic interpretation and LLM hate-speech calibration as the principal areas requiring further refinement.

Future work should proceed in two complementary directions. At the pilot-resource level, expert adjudication can strengthen two-of-three human decisions and unresolved multiclass cases while preserving the original disagreement information, and the annotation guidelines can be refined particularly for target inference and profanity function. At the scaling level, the human-validated pilot provides a basis for recalibrating the LLM first-pass pipeline before applying it to the larger collected corpus, with selective human review and independently annotated validation subsets used to monitor annotation quality. This would enable the present pilot resource and annotation methodology to serve as the basis for a substantially larger Serbian harmful-language dataset and for subsequent training and evaluation of target-aware and explainable detection systems.

9. Limitations

Several limitations should be considered when interpreting the results.

First, the study is based on an intentionally enriched Reddit sample. Lexicon-guided retrieval and phenomenon-oriented pilot selection were used to increase the presence of relevant and difficult cases, so the reported label distributions should not be interpreted as population-level prevalence estimates. In addition, linguistic and interactional patterns may differ across other platforms and genres.

Second, the human reference is based on majority agreement among three independent annotators rather than expert adjudication. Two-of-three decisions therefore preserve genuine disagreement, and the resulting labels are best understood as a human-consensus reference rather than an adjudicated gold standard.

Third, annotation reliability varies across dimensions. Explicit phenomena such as profanity presence are comparatively reproducible, while target-related and pragmatic distinctions are more difficult. These differences are part of the empirical findings of the study and identify areas in which further guideline refinement may be useful.

Fourth, annotations were made primarily from isolated comment text rather than complete discussion threads. Wider conversational context could affect the interpretation of irony, implicit targets, reported speech, and other borderline cases.

Finally, the automated evaluation concerns a single structured LLM first-pass configuration. Different models, prompts, or calibration strategies may produce different error profiles, and the observed hate-speech overclassification should therefore be interpreted as a property of the evaluated pipeline rather than of LLM-based annotation in general.

References

- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. [Automated Hate Speech Detection and the Problem of Offensive Language](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–517.
- Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth M. Belding. 2018. [Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media](#). In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, pages 42–51. AAAI Press.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks](#). *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Danka Jokić, Ranka Stanković, Cvetana Krstev, and Branislava Šandrih Todorović. 2021. [A Twitter Corpus and Lexicon for Abusive Speech Detection in Serbian](#). In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.
- Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. [The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English](#). In *Text, Speech, and Dialogue*, volume 11697 of *Lecture Notes in Computer Science*, pages 103–114. Springer.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Bie-mann, Pawan Goyal, and Animesh Mukherjee. 2021. [HateX-plain: A Benchmark Dataset for Explainable Hate Speech Detec-](#)

- tion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.
- Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott A. Hale, and Helen Margetts. 2019. [Challenges and Frontiers in Abusive Content Detection](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.
- Zeera Waseem, Thomas Davidson, Dana Warmusley, and Ingmar Weber. 2017. [Understanding Abuse: A Typology of Abusive Language Detection Subtasks](#). In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. [Predicting the Type and Target of Offensive Posts in Social Media](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1415–1420.
- Ema N. Živković Nikolić. 2024. [Impoliteness in Online Discourse: Analysis of Reader Comments on Serbian News Sites](#). *Philologia Mediana*, 16(16):655–671.

Ka višedimenzionalnoj anotaciji usmerenog govora mržnje u srpskom jeziku

Miloš Košprdić, Isidora Dekić, Tijana Radović

Sažetak

Detekcija govora mržnje predstavlja složen zadatak jer se uvredljivi govor (engl. abusive speech), psovke, omalovažavanje zasnovano na identitetu i hostilnost usmerena prema društvenim grupama često prepliću, iako ne predstavljaju iste komunikativne pojave. Ovo razgraničenje posebno je značajno u srpskom onlajn diskursu, u kojem uvredljivi i identitetski obojeni izrazi mogu imati različite pragmatičke funkcije i ne moraju nužno predstavljati govor mržnje. U radu predstavljamo višedimenzionalnu anotacionu šemu i prateće smernice za govor mržnje i uvredljivi govor u srpskom jeziku, kojima se eksplicitno modeluju struktura i vrsta mete, omalovažavanje zasnovano na identitetu, funkcija psovki, citirani i preneseni uvredljivi govor i autorski stav. Primenom leksički i fenomenološki usmerenog prikupljanja konstruisan je veći korpus komentara sa platforme Reddit, iz kojeg je izdvojen namerno obogaćen pilot skup od 500 komentara. Pilot skup je najpre anotiran strukturisanim postupkom zasnovanim na velikom jezičkom modelu, a zatim nezavisno od strane više ljudskih anotatora, što je omogućilo procenu pouzdanosti anotacione šeme, formiranje ljudskog konsenzusa i neposredno poređenje odluka modela i ljudskih anotatora. Rezultati potvrđuju jasnu razliku između uvredljivog govora, omalovažavanja zasnovanog na identitetu i govora mržnje, dok samo prisustvo psovki ima malu diskriminativnu vrednost. Pragmatičke odluke povezane sa funkcijom psovki, tumačenjem mete i autorskim stavom pokazale su se zahtevnijim od prepoznavanja eksplicitnih jezičkih pojava, dok je veliki jezički model pokazao sistematsku sklonost ka preširokom dodeljivanju oznake govora mržnje uvredljivom i identitetski obojenom sadržaju. Rad pruža operacionalizovanu anotacionu šemu, prateće smernice i ljudski validiran pilot resurs za srpski jezik, kao i empirijsku osnovu za unapređenje anotacije podržane velikim jezičkim modelima i njeno dalje skaliranje na veće kolekcije podataka.

Ključne reči: govor mržnje, uvredljivi govor, višedimenzionalna anotacija, obrada prirodnog jezika, srpski jezik, ljudska anotacija, anotacija podržana velikim jezičkim modelima

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>

