ParsHate: A Benchmark Dataset for Hate and Target Detection in Persian

Anonymous EMNLP submission

Abstract

We introduce ParsHate, a manually annotated dataset of 10,000 Persian tweets spanning 2013–2022, representing the first decade-long benchmark for hate speech detection in Persian. The dataset contains 31% hateful content and supports both hate detection and multi-label fine-grained target identification across seven structured target categories. ParsHate also distinguishes explicit and implicit hate, marks explicit and implicit targets, and provides span-level rationales. Data collection combines random and score-stratified temporal sampling to reduce keyword-driven bias while preserving natural label distributions. Applying SOTA models for Persian hate-speech detection on ParsHate shows moderate performance (79% F1), especially with samples from earlier years, and low performance with target identification (25.5% macro-F1). This emphasizes the diverse sampling of hate speech in ParsHate and its challenging nature that requires more advanced methods for better performance. Dataset is made publicly available.

1 Introduction

Hate speech detection has become a central task in Natural Language Processing due to its social impact and the growth of user-generated content (Bokaei et al., 2025). Over the past decade, many annotated datasets have supported automatic hate speech detection, especially in English and other high-resource languages. These datasets typically label hateful and non-hateful content, and may include hate targets or span-level rationales identifying the evidence behind the label (Davidson et al., 2017; Mathew et al., 2021; Kennedy et al., 2022). Some also distinguish explicit from implicit hate, reflecting the complexity of modelling harmful language online (ElSherief et al., 2021). Recent datasets further provide structured target taxonomies and span-level rationales, enabling analysis of both hate expressions and the identity groups involved (Mathew et al., 2021; Kennedy et al., 2022; ElSherief et al., 2021; Davidson et al., 2017). These schemes provide richer supervision and support more detailed model evaluation.

Despite these advances, Persian remains comparatively under-resourced. To our knowledge, only a limited number of datasets have been developed for Persian hate speech detection, including Pars-Off (Ataei et al., 2022), PHATE (Delbari et al., 2024), PHICAD (Davardoust et al., 2024), and Pars-HAO (Sheykhlan et al., 2023). Although these resources have significantly advanced research in Persian harmful language modeling, they exhibit several structural limitations. Pars-Off relies heavily on keyword-driven sampling and tweets extracted from politically sensitive accounts, which can lead to lexical shortcut learning and under-representation of implicit forms of harassment. PHATE introduces span-level rationales and transformer-based baselines, relies partially on keyword-based candidate selection and covers a narrow temporal window. In addition, its target annotations are provided in free text and have not been subject to a normalized schema. PHICAD substantially increases the scale of Persian harmful content data, but does not provide a target taxonomy or distinctions between explicit and implicit hate.

These limitations make it difficult to systematically examine how hate expression evolves across time, how hostility is conveyed implicitly, and how hate targets are represented in Persian online discourse. To address these limitations, we introduce ParsHate, a decade-spanning benchmark dataset of 10,000 manually annotated Persian tweets covering the period from 2013 to 2022, where 31% of tweets annotated to contain hate-speech. The dataset is constructed using a hybrid temporal sampling strategy that combines fully random sampling with score-stratified sampling based on predictions from

a Persian toxicity classifier (Bokaei et al., 2025). This design mitigates keyword-driven sampling bias while ensuring coverage across varying levels of hate without artificially enforcing label balance.

ParsHate supports hate detection and multi-class target identification across seven structured target categories—Gender, Politics, Religion, Nationality, Occupation, Ethnicity, and Other—with fine-grained subcategories. Following prior work on target-aware annotations for implicit hatred (Bokaei et al., 2025), the dataset distinguishes explicit from implicit hate and categorizes implicit expression strategies. It also annotates whether the target itself is expressed explicitly or implicitly and provides span-level rationales highlighting the textual evidence supporting each annotation. These annotations enable further investigation of implicit hate and support the development of more robust multi-label target prediction models. Using ParsHate, we establish benchmark experiments under single year training, integrated multi-year training, and cross-temporal transfer settings to examine how detection performance varies across time and how distributional shifts affect model stability. Our contributions are as follows:

- Longitudinal Temporal Benchmark: We introduce ParsHate, the first decade-spanning (2013–2022) Persian hate speech dataset and establish cross-year evaluation benchmarks to measure temporal distribution shifts.
- Structured Target Modeling: We provide fine-grained multi-class target annotations with explicit identity dimensions and explicit/implicit target marking.
- Implicit–Explicit Hate Annotation: We annotate hate as explicit or implicit, label the strategy used, and provide span-level rationales to support interpretability research.

2 Related Work

Research on hate speech detection has produced numerous annotated datasets across languages and platforms. We organise this section around the four dimensions that motivate ParsHate: target taxonomies, implicit hate, span-level rationales, and temporal coverage. This makes the gaps in existing Persian resources explicit and situates ParsHate within the broader dataset landscape.

Target Taxonomies: In English, Mathew et al. (2021) proposed HateXplain, annotated with class labels and target categories, and Toraman et al. (2022) constructed English and Turkish datasets across five target categories. Kennedy et al. (2022) introduced the Gab Hate Corpus, a 27K-post dataset with identity-based target categories. In Persian, Pars-OFF (Ataei et al., 2022) provides coarse target types, but not fine-grained identity categories. PHATE (Delbari et al., 2024) provides free-text target annotations rather than a normalised taxonomy. Pars-HAO (Sheykhlan et al., 2023) and PHICAD (Davardoust et al., 2024) do not provide structured target schemas.

Implicit and Coded Hate: A growing body of work in English addresses implicit and coded hate, where hostility appears without overt slurs or profanity. ElSherief et al. (2021) introduced a taxonomy of implicit hate with fine-grained categories and target annotations, highlighting the difficulty of modelling indirect hostility. Hartvigsen et al. (2022) proposed TOXIGEN, a large-scale machine-generated dataset balanced across identity groups to improve robustness to subtle toxicity. Kennedy et al. (2022) additionally annotate an implicitness label in the Gab Hate Corpus, although they do not distinguish whether the target itself is explicitly named or indirectly referenced. Prior work highlights the difficulty of modelling implicit hatred in Persian (Bokaei et al., 2025; Delbari et al., 2024), yet no existing Persian dataset annotates implicit hate strategies or distinguishes implicit targets.

Span-Level Rationales for Interpretability: Span-level rationales support interpretability in English benchmarks such as HateXplain (Mathew et al., 2021), the Gab Hate Corpus (Kennedy et al., 2022), and Latent Hatred (ElSherief et al., 2021). In Persian, only PHATE (Delbari et al., 2024) provides them.

Temporal Coverage and Sampling: Studies in high-resource languages show that hateful discourse varies over time (Mathew et al., 2020; Garland et al., 2022), motivating datasets that support longitudinal analysis. Existing Persian resources, however, cover limited temporal spans: PHATE (Delbari et al., 2024) covers 2020–2023, Pars-OFF (Ataei et al., 2022) covers 2015–2020, and the temporal windows of Pars-HAO (Sheykhlan et al., 2023) and PHICAD (Davardoust et al., 2024) are not reported. These datasets also rely on keyword- or profile-based sampling, which introduces lexical bias and

Dataset	Source	Period	Size	% hate	implicit hate	target	implicit target
Pars-OFF (Ataei et al., 2022)	Tweets	2015-2020	10000	30% (OFF)	X	X	X
PHATE (Delbari et al., 2024)	Tweets	2020-2023	7000	23%	X	✓	X
Pars-HAO (Sheykhlan et al., 2023)	Tweets	unknown	8000	9%	X	X	X
PHICAD (Davardoust et al., 2024)	Instagram	unknown	30000	41% (HOF)	X	X	X
ParsHate (ours)	Tweets	2013-2022	10,000	31%	✓	✓	✓

Table 1: How ParsHate compares to existing Persian hate-speech datasets

under-represents implicit forms of hate.

Positioning of ParsHate: Other widely used English benchmarks, including (Tonneau et al., 2024, 2025; Röttger et al., 2021; Salminen et al., 2018; Calabrese et al., 2025; Ahn et al., 2024; Shen et al., 2025; Ghorbanpour et al., 2025), provide large-scale annotations, target-aware labels, or diagnostic evaluation suites for hate detection. Across the four dimensions above, no existing Persian resource combines structured target taxonomies, implicit-hate strategy annotations, span-level rationales, and longitudinal coverage. ParsHate addresses these gaps by providing a seven-category target taxonomy with fine-grained subcategories, annotating implicit hate with three rhetorical strategies and implicit target marking, pairing span-level rationales with strategy labels, and covering a decade of Persian hate speech (2013–2022). It also uses score-stratified temporal sampling to reduce keyword-driven bias while preserving natural label distributions. Table 1 summarises these contributions.

3 ParsHate: Dataset Construction

In this section, we describe the construction of ParsHate, a Persian hate speech dataset collected from Twitter. We detail our data retrieval strategy, annotation design, target taxonomy development, and quality control procedures, highlighting the methodological choices made to ensure linguistic diversity, target coverage, and reliable labeling.

3.1 Data Collection

To construct ParsHate, we extracted Persian tweets from a Twitter archive containing a 1% daily sample of the Twitter stream between 2013 and 2022. The archive included approximately 47 million Persian tweets. After removing retweets, duplicates, tweets shorter than five tokens, and tweets containing only URLs, mentions, or hashtags, the corpus was reduced to 29 million tweets. This filtered pool was used for the hybrid temporal sampling procedure described below.

To enable systematic sampling across different

Year	tweet	Hate	Hate Implicit?	Implicit Strategy	Hate target	Target Implicit?
2022	مردم ما دیگه گول سلبریتی نمیخورن گاییدین مارو با چسناله! عامل اصلی ناامیدی بین مردممونن! همین ملیجکا! Our people don't fall for celebrities anymore. You've screwed us over with your whining! These same little lapdogs are the main cause of despair among our people!	1	0	0	Occupation	0
2019	نه پس ولایت مطلقه وحوشی مثل ترامپ و طرفداراش خوبه 😊 No then—absolute rule by beasts like Trump and his supporters is good? 😊	1	1	Figurative Language	Politics - Foreign	0
2017	همون آقایی که سال ۹۴ بخاطر اخم جان کری نمیخواست کشتی کمک‌های انسان‌دوستانه ایران بره یمن، حالا میگه دیپلماسی فدای میدان شده! خودش و طرفداراش مملکت رو به خاک سیاه کشیدن! The same guy who in 2015 didn't want Iran's humanitarian aid ship to go to Yemen because John Kerry frowned is now saying diplomacy was sacrificed for the battlefield! He and his supporters have dragged the country into ruin.	1	1	Reference	Politics-Domestic	1
2013	از همجنس بازای کثیف نظر نخواستیم. خب مشخصه جماعت هرزه همجنسباز چیزی جز این حرفا بلد نیستن . We didn't ask for the opinion of filthy homosexuals. Obviously that group of immoral homosexuals doesn't know how to say anything other than this kind of nonsense	1	0	0	Ethnicity	0

Table 2: Dataset instances with corresponding labels.

levels of hate likelihood, we applied a Persian hate classifier based on fine-tuned Llama 3 (Bokaei et al., 2025), which showed the SOTA performance on existing Persian hate-speech datasets. The model was trained onhate speech datasets in Persian, Arabic and Indonesian, showing a strong performance in prior evaluations. Importantly, none of the datasets used to train this classifier overlap with the tweets included in ParsHate. The model was applied to the filtered corpus to assign probability scores indicating the likelihood of hateful content. These scores were used solely for stratified sampling and not for automatic labeling.

For each year between 2013 and 2022, we sampled 1,000 tweets using a hybrid strategy to balance representativeness and hate intensity coverage. Specifically, 500 tweets were selected randomly from each year to capture naturally occurring discourse. The remaining 500 tweets were selected using classifier-based stratification, with 50 tweets randomly sampled from each probability interval (0.0–0.1, . . . , 0.9–1.0), where the probability value indicates the likelihood of a tweet to contain hate-speech. This design ensures inclusion of neutral content, borderline cases, and high-confidence hate instances while mitigating lexical bias and preserving temporal diversity.

3.2 Annotation Scheme and Guideline

ParsHate employs a structured annotation framework designed to capture hate presence,

rhetorical realization, and identity targets.

Hate Definition: Following (Waseem and Hovy, 2016), we define hate as content that expresses or encourages hostility toward groups or individuals based on their association with identity-related categories. This includes negative stereotyping, discrimination, dehumanization, humiliation, or encouragement of harm directed at social or political groups. Each tweet is assigned a binary hate-speech label: hate or non-hate.

Target Taxonomy: The target taxonomy is inspired by identity categories used in the Gab Hate Corpus (Kennedy et al., 2022), including gender, political identification, religion, nationality, ethnicity, and other. We extend this framework by adding occupation to reflect identity-based targeting patterns observed in Persian discourse in previous work (Delbari et al., 2024). Target annotation allows multi-label assignment when multiple groups are attacked within a single tweet.

Within high-level categories, annotators specified fine-grained subtypes when applicable, following a closed set finalised through the pilot annotation phase and inter-annotator discussion (Appendix A.5). Gender distinguishes between *men* and *women*. Religion distinguishes *Islam* from *other religious identities*, reflecting the empirical predominance of Islam-targeted hate in Persian discourse (Delbari et al., 2024). Politics distinguishes *domestic* from *foreign* political actors. Nationality distinguishes *Iranian*, *Afghan*, *Arab*, *Western*, and *other*. Occupation, Ethnicity, and the catch-all Other category are not further subdivided. When the target of hate is implied without being explicitly mentioned, it is annotated as an *implicit target*.

Implicit Hate: Following prior work that distinguishes between explicit and implicit hate (Kennedy et al., 2022), we annotate whether the hateful content is expressed explicitly or implicitly. Implicit hate refers to hostility conveyed without direct slurs or overt statements of hatred, often relying on cultural knowledge, metaphor, sarcasm, insinuation, or exclusionary logic. Based on a pilot annotation phase (250 tweets), we identified three primary strategies of implicit hate: (i) cultural or historical reference, (ii) sarcasm or figurative language, (iii) other indirect rhetorical techniques. Annotators were allowed to assign more than one strategy when multiple rhetorical mechanisms co-occurred within the same tweet. We additionally annotate whether the target itself is explicit or implicit, distinguishing between cases where the attacked group is directly named versus indirectly invoked (Kennedy et al., 2022)

Span-Level Annotation: For hate tweets, annotators highlighted the shortest text segment(s) justifying the label. For explicit hate, this was typically overt hate lexicon; for implicit hate, it was the shortest rhetorical device anchoring hostility, such as metaphor, sarcasm, or cultural/historical reference, or the minimal hostile clause when no device was lexically isolable. This follows rationale conventions in HateXplain (Mathew et al., 2021), the Gab Hate Corpus (Kennedy et al., 2022), and Latent Hatred (ElSherief et al., 2021). A 250-tweet pilot aligned annotators on implicit-span cases. Span selections were converted into character-level binary labels. A character was retained as part of the final span if at least two annotators marked it; otherwise, it was discarded. Final spans were defined as maximal contiguous sequences of majority-selected characters.

Annotation Procedure and Agreement: All 10,000 tweets were independently annotated by three native Persian speakers following detailed written guidelines (Annotator demographics and annotation guidelines are reported in the Appendix.). A pilot phase was conducted to refine definitions, particularly for implicit hate and implicit targeting. Following Röttger et al. (2022), ParsHate uses a *prescriptive* annotation paradigm, with agreement treated as a quality target and disagreements resolved into single instance-level labels through majority voting and discussion-based adjudication. Inter-annotator agreement was measured using Fleiss' κ. Agreement scores were κ = 0.69 for hate identification, κ = 0.60 for target category, κ = 0.51 for implicit hate, and κ = 0.58 for implicit target. For span annotations, agreement was computed at the character level, yielding κ = 0.55 overall, with κ = 0.64 for explicit spans and κ = 0.44 for implicit spans. Span disagreements were resolved through majority-character aggregation and discussion-based adjudication.

3.3 Disagreement Analysis and Resolution

3.4 Disagreement Analysis and Resolution

Beyond aggregate agreement, we analyse disagreement patterns across the three implicit-hate strategies. Per-strategy Fleiss' κ values invert the

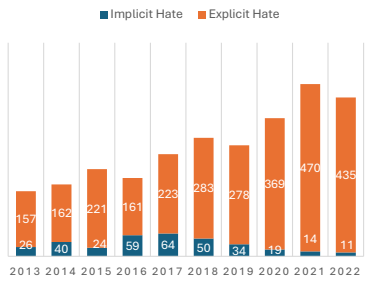

Figure 1: Hate Samples Distribution Across years.

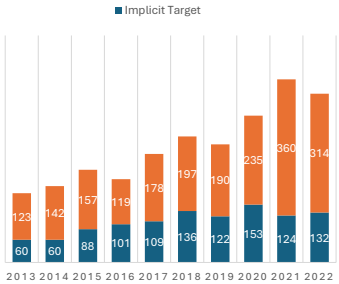

Figure 2: Implicit-Explicit Hate Target Distribution Across years.

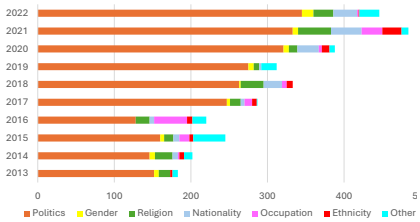

Figure 3: Hate Target Distribution Across Years.

frequency ordering: cultural/historical reference, the most common strategy (48% of implicit-hate tweets), is also the most disagreement-prone ($\kappa = 0.43$), followed by sarcasm/figurative language ($\kappa = 0.55$) and other indirect strategies ($\kappa = 0.58$). Cultural-reference disagreements mainly reflected unequal recognition of political or historical allusions; sarcasm disagreements concerned whether figurative content was *hate* or merely *critical*. Implicit-hate rationale spans are also longer and more clause-level than explicit ones, and contain more named entities (43% vs. 13%), reflecting references to politicians and historical figures. Roughly 16% of implicit-hate and 6% of explicit-hate tweets required adjudication beyond majority voting, resolved through guideline-based discussion with unresolved ties defaulting to the majority label. Full per-strategy statistics appear in Appendix A.4.

3.5 Dataset Statistics and Analysis

ParsHate consists of 10,000 Persian tweets collected between 2013 and 2022. Hate content constitutes 31% of the dataset. As represented in Figure 1, temporal analysis reveals a steady increase in hate expression over the examined period. The proportion of tweets expressing hate rises from 5.90% in the earliest year to a peak of 15.61% in later years. With respect to explicitness, 7% of hate tweets are annotated as implicit, suggesting that most hateful expressions in this corpus are overt. However, target expression exhibits a different pattern: 35% of hate tweets contain implicit targets, where the attacked group is not directly named. Figure 2 represents the distribution of implicitness for both hate and target over the years. Target distribution, which is heavily skewed toward politics, which accounts for 76.45% of hate instances. Religious (6.42%), nationality (4.90%), occupation (3.45%), ethnicity (2.29%), gender (1.94%), and other categories (4.55%). Notably, 39% of political hate instances involve implicit target references. Span-level analysis further shows that hate spans cover, on average, 18% of tweet characters. Figure 3 represents the distribution of targets over time.

4 Analysis of ParsHate

In this section, we analyze the temporal and structural properties of ParsHate. Specifically, we examine (i) temporal robustness of hate detection models across years; (ii) lexical variability across time using type–token ratio (TTR) and named entity masking; and (iii) multi-label target prediction behavior under class imbalance. Together, these analyses provide insight into the challenges posed by the dataset for hate and multi-label target prediction models.

4.1 Hate-Speech Detection over Time

To analyze temporal variation in ParsHate, we evaluate the generalization capacity of hate detection models across years (2013–2022). For each year, the data is randomly split into 80% training and 20% testing, and this split is kept fixed across all experiments to ensure comparability. We fine-tuned both LLaMA-3 and Gemma-3; as Table 3 shows, the two models yield nearly identical performance across all settings. All reported results use macro-F1 and are averaged over three random seeds. We consider three training strategies to analyze temporal robustness: (i) a PHATE-trained state-of-the-art (SOTA) LLaMA-3 model, trained using the original PHATE training split (2020–early 2023) (Delbari et al., 2024) and evaluated on each ParsHate year. PHATE and ParsHate share no overlapping tweets; for schema compatibility, although PHATE includes hateful, violent, and vulgar labels, we use only the *hateful* label to align with ParsHate's binary hate/non-hate scheme;

(ii) integrated training on 80% of data from all ParsHate years (*all-years*), followed by evaluation on each year separately; and (iii) year-specific training, where models are trained and tested within the same year.

Table 3 presents the performance of all experiments. The integrated *all-years* training strategy achieves the highest overall performance (78.5 macro-F1), consistently outperforming the PHATE-trained SOTA model (74.2) and year-specific models (62.9–68.0 across architectures). This suggests that temporally diverse training data improves robustness and mitigates year-specific distribution. The PHATE-trained SOTA model exhibits strong performance on temporally proximate years (2020–2022), reaching macro-F1 above 88, but performance declines substantially in earlier years (2013–2015), where scores fall to the mid-50s to mid-60s range. This degradation can reflect a temporal distribution shift, indicating that models trained on recent discourse struggle to generalize to earlier contexts. In contrast, integrated training across all years produces more stable performance across the temporal spectrum, reducing early-year degradation and improving average robustness. Year-specific training yields lower overall performance, suggesting that limited yearly data constrains generalization capacity.

4.2 Lexical Variability and NE Influence

Given the observed performance degradation in the early years, we next examine whether lexical variability contributes to the shift in temporal distribution in ParsHate. To quantify lexical diversity, we compute the Type-Token Ratio (TTR) (McEnery and Hardie, 2011), defined as the ratio of unique words (types) to total words (tokens) in a given corpus. Before computation, tweets are normalized and tokenized using Hazm, a Persian tokenizer [1] which has demonstrated strong performance for Persian text processing (Kamali et al., 2022). We observed that the early years (2013–2015) exhibit the highest lexical diversity, followed by a gradual decline over time with minor fluctuations. Later years (2019–2022) show comparatively lower TTR values, indicating increased lexical stabilization in recent discourse. In the next step, to investigate whether named entities contribute to lexical variability, we perform an additional analysis by masking named entities using ParsBERT NER- as demonstrated in the Persian NER task, (Farahani et al., 2021), and recomputing TTR. Although lexical diversity decreases after masking, early years remain consistently more diverse than later years. Notably, the strongest masking effect occurs in the earliest years (2013–2014), suggesting that named entities can contribute more to lexical variability during that period. From 2016 onward, the reduction in TTR after masking becomes smaller and stabilizes. Figure 4 in the Appendix presents Lexical variability across years before and after masking. These findings indicate that early-year discourse is characterized by greater topical and lexical variation, partially driven by named entities. The higher lexical diversity in early years provides a plausible explanation for the reduced performance of temporally distant models observed in the previous subsection, as models trained on later, more stabilized discourse encounter broader lexical variation when applied to earlier data.

4.3 Multi-label Target Prediction

We formulate target prediction as a multi-label classification task: given a hate tweet, the model predicts which of the seven target categories (Gender, Politics, Religion, Nationality, Occupation, Ethnicity, Other) are attacked, allowing more than one label per tweet when multiple groups are targeted. This formulation follows the protocols of Zampieri et al. (2019) and Lu et al. (2023). As shown in Figure 3, political targets account for 76.45% of hate instances, while other categories appear substantially less frequently. This imbalance creates a realistic but challenging setting for multi-target modeling. We evaluate fine-tuned LLaMA-3 and GPT-5 for multi-target classification. As these models achieved the strongest performance in earlier hate detection experiments, we focus the multi-label target prediction analysis on them. We assess robustness by comparing performance before and after NE masking.

As represented in Table 5 LLaMA-3 demonstrates high precision for the Politics class (F1 = 0.89) but suffers from severe recall degradation on low-frequency targets such as Occupation and Ethnicity, resulting in near-zero F1 scores. In contrast, GPT-5 consistently achieves substantially higher recall for minority targets

[1] https://pypi.org/project/hazm/0.9.1/

Year	Zero-shot GPT-5			SOTA (Bokaei et al., 2025)			Fine-tuned *year-by-year* ParsBERT			LLaMA 3			Gemma 3			Fine-tuned *all-years* LLaMA 3			Gemma 3		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
2013	54.6	**66.2**	59.9	**65.4**	51.7	57.9	X	X	X	42.5	40.3	41.4	43.8	45.6	44.7	64.3	63.8	**64.0**	63.2	60.4	61.7
2014	51.4	62.7	56.4	**68.2**	49.5	57.3	X	X	X	48.6	42.4	45.3	55.7	38.6	45.5	67.5	**66.3**	**66.9**	67.3	65.6	66.4
2015	50.8	50.4	50.6	69.6	61.3	65.2	X	X	X	63.7	59.8	61.7	60.4	56.9	58.6	79.2	75.5	77.3	**80.4**	**78.1**	**79.2**
2016	66.7	62.3	64.4	**72.5**	74.6	73.5	X	X	X	60.5	51.9	55.9	54.6	56.2	55.4	71.6	**79.3**	**75.2**	71.4	78.3	74.7
2017	58.6	56.9	57.7	73.8	65.2	69.2	61.3	52.4	57.2	70.6	62.8	66.5	60.9	68.4	64.4	**81.4**	79.8	80.6	80.1	**82.3**	**81.1**
2018	68.9	72.3	70.6	75.4	77.8	76.6	64.3	62.1	63.2	70.8	68.4	69.6	74.2	72.9	73.5	78.6	80.4	**79.5**	80.5	**82.1**	**81.3**
2019	76.8	74.7	75.7	82.6	**88.3**	85.3	59.1	63.3	61.1	62.3	64.5	63.4	66.9	64.8	65.8	**90.4**	86.6	**88.4**	89.3	85.6	87.4
2020	**88.5**	76.9	**82.3**	78.7	**84.1**	81.3	65.4	69.5	67.4	67.8	71.6	69.6	69.5	73.4	71.4	77.5	81.7	79.5	79.3	83.3	81.4
2021	67.9	71.5	69.7	**93.4**	**88.6**	**90.9**	78.4	79.1	78.7	85.4	77.6	81.3	82.6	84.9	83.7	85.7	87.6	86.6	85.0	81.2	83.1
2022	**93.6**	87.8	**90.6**	91.3	85.7	88.4	79.0	79.1	79.1	76.4	82.5	79.3	79.3	75.8	77.5	89.5	**91.2**	90.3	90.4	89.1	89.7
All	67.8	68.2	68.0	77.1	72.7	74.8	68.0	67.5	67.7	64.9	62.2	63.5	64.8	63.8	64.3	**78.6**	**79.2**	**79.0**	**78.6**	77.7	78.2

Table 3: Precision (P), Recall (R), and F1 (F) scores across years and models.

Hate-type	LLaMA-3 P	LLaMA-3 R	LLaMA-3 F1	Gemma-3 P	Gemma-3 R	Gemma-3 F1
Implicit	5.3	3.7	4.4	**8.3**	**5.0**	**6.2**
Explicit	**84.6**	**84.4**	**84.5**	84.3	83.1	83.7
Overall	**78.6**	**79.2**	**79.0**	**78.6**	77.7	78.2

Table 4: Models' performance according to the explicitness of hate. Models are trained on all years.

Target	Unmasked GPT-5 P	Unmasked GPT-5 R	Unmasked GPT-5 F1	Unmasked LLaMA-3 P	Unmasked LLaMA-3 R	Unmasked LLaMA-3 F1	Masked GPT-5 P	Masked GPT-5 R	Masked GPT-5 F1	Masked LLaMA-3 P	Masked LLaMA-3 R	Masked LLaMA-3 F1
Gender	**22.4**	**80.3**	**35.1**	18.2	13.1	15.2	21.3	**80.6**	33.7	11.6	7.1	8.8
Politics	96.3	32.7	48.8	84.6	96.2	**90**	**97.4**	27.6	43.0	82.3	**98.5**	**89.6**
Religion	14.5	**81.2**	24.6	65.1	32.7	**43.6**	14.7	**81.4**	24.6	**71.5**	12.4	21.1
Nationality	7.3	**60.7**	13.4	43.4	8.2	13.8	8.4	**60.5**	14.8	**57.2**	10.3	**17.5**
Occupation	5.5	60.4	10.1	20.7	2.1	3.9	8.2	**65.7**	14.6	**33.3**	3.2	**33**
Ethnicity	4.7	60.1	8.8	**15.3**	2.6	4.4	6.6	**65.2**	**12.1**	14.2	2.3	4.0
Other	30.5	22.3	25.8	**46.3**	6.3	11.1	25.2	**33.5**	**28.7**	33.4	2.7	5.1
Macro avg	25.5	56.6	23.8	42.0	23.0	26.0	25.9	**59.1**	24.4	**42.8**	19.5	**27.1**

Table 5: Per-target performance comparison across models.

(often above 0.60), leading to improved F1 scores for rare categories along with low precision. To assess whether models rely on named entities for target prediction, we repeat the evaluation after masking named entities using ParsBERT (Farahani et al., 2021). Masking reduces overall performance for both models as represented in Table 5 but preserves the fundamental behavioral contrast: LLaMA-3 remains dominant on Politics with high precision, while GPT-5 maintains broader recall across minority targets. Overall, these findings highlight that ParsHate presents a structurally realistic yet challenging target distribution, where dominant political hate coexists with sparse identity-based categories. The contrasting behaviors of LLaMA-3 and GPT-5 further illustrate the tradeoff between precision and recall in low-resource target prediction settings.

4.4 Error Analysis

To better understand model limitations under temporal diversity and structural imbalance, we conduct a detailed error analysis of the best-performing *all-years* model (LLaMA-3-FT) and compare its behavior with GPT-5 as the next strong model. LLaMA-3-FT and Gemma-3-FT achieve nearly identical overall performance under the *all-years* setting (Table 3); we therefore focus the error analysis on LLaMA-3-FT as a representative case. We focus separately on hate detection and multi-label target prediction. Table 8 in the Appendix represents some misclassification samples on these models.

Hate Detection Error: Although implicit hate constitutes only 7% of hate instances in the dataset, it accounts for a large share of false negatives. In the *all-years* model, implicit hate represents approximately 31% of missed hate instances. Breaking down implicit errors by strategy reveals that cultural or historical references are the most challenging category, followed by sarcasm or figurative language, while the "other indirect" category contributes comparatively fewer errors. This suggests that instances requiring background knowledge and contextual interpretation are more likely to remain undetected than cases containing overt sentiment cues. In contrast, GPT-5 exhibits a lower concentration of implicit false negatives, with implicit hate accounting for approximately 22% of its missed hate instances. However, GPT-5 introduces a higher number of false positives in figurative contexts. Overall, implicit rhetorical strategies—particularly those grounded in cultural or historical reference—constitute a large proportion of misclassification cases. even under temporally integrated training.

Explicit vs. Implicit Performance Breakdown. Using ParsHate's explicit/implicit annotations, we

evaluate the two strongest fine-tuned models (LLaMA-3-FT and Gemma-3-FT) on explicit- and implicit-hate test subsets separately (Table 4). Both perform well on explicit hate, with average F1 = 84 for both models, but collapse on implicit hate, with F1 = 3 and F1 = 6, respectively. For LLaMA-3-FT, 83% of implicit-hate errors are predicted as non-hate, 12% as hate with the wrong target, usually *politics*, and 5% are partial-credit failures in multi-label cases. GPT-5 partly reduces this recall collapse and implicit hate accounts for 22% of its missed hate cases versus 31% for LLaMA-3-FT but over-flags neutral metaphor as hate in figurative contexts. Thus, neither model resolves implicit hate: LLaMA-3-FT mainly misses it, while GPT-5 improves recall at a precision cost. This suggests that background knowledge must be paired with the ability to distinguish hateful from neutral cultural references. Representative cases are provided in Appendix Table 7. Full per-strategy implicit-hate analysis is provided in Appendix A.1. Span-detection performance are analysed in Appendix A.2.

Multi-label Target Prediction Error: multi-label target prediction errors reveal that while implicit targets account for 35% of hate instances in the dataset, they represent approximately 58% of target misclassifications in the *all-years* model. This confirms that target identification under indirect reference remains substantially more difficult than detecting the presence of hostility itself. LLaMA-3-FT exhibits that misclassified minority targets (e.g., occupation, ethnicity, gender) are frequently predicted as Politics, the majority class. This minority collapse reflects the model's reliance on dominant target distributions when explicit identity cues are weak or absent. GPT-5 demonstrates a contrasting pattern. While it achieves higher recall for minority targets, it frequently over-assigns target categories, leading to lower precision. In several cases, GPT-5 predicts multiple plausible targets even when only one is supported by the text.

Qualitative Analysis of Annotator Rationales on Misclassifications. To complement the quantitative error breakdown, we examine annotator-selected rationale spans for misclassified implicit-hate and implicit-target cases in the *all-years* model. Two representative cases illustrate distinct failure modes; the full set appears in Appendix Table 6. In the first case, a tweet ending with "what colour rope do you prefer, you British-bred mullahs?" is annotated as implicit hate, with reference and sarcasm as rhetorical strategies. The hate is a culturally specific euphemism for execution, framed as a mock offer of choice. No token is hateful in isolation; recognising it requires knowledge that "offering a rope" encodes a call for hanging. The *all-years* model predicts non-hate, illustrating the dominant implicit-hate failure mode: cultural or historical rationales convey hate only with external knowledge.The second case shows a failure mode tied to implicit targets. In a tweet mourning *Mahsa Amini* a widely publicised case and focal point of political protest in Iran in 2022, annotators highlighted spans including "bastards" and "what did you do to this girl," reflecting hostility toward the regime. The model detects hate but assigns the target as *gender*, based on the surface mention of "this girl." Here the hate is explicit, but the target is implicit and discourse-level: the addressees are state actors, not a gender group. This shows why implicit-target identification requires reasoning beyond local lexical signals. Across the full set, rationales show where cultural, historical, or rhetorical knowledge is needed and where the model lacks signal. They serve both as supervision signals and as diagnostic evidence of the background knowledge current models fail to capture. Implicitness is a central source of modelling difficulty in ParsHate: even strong models struggle when hate is indirect or targets are underrepresented. Together with temporal variation, lexical diversity, and class imbalance, it shapes the dataset's core challenges.

5 Conclusion

We introduce ParsHate, the first decade-spanning (2013–2022) Persian benchmark for hate and multi-label target prediction, with 10,000 manually annotated tweets, a structured target taxonomy, explicit/implicit distinctions, strategy categories, and span-level rationales. Score-stratified temporal sampling preserves natural label distributions while reducing keyword-driven bias. Evaluation reveals both temporal distribution shift across years and structural collapse on implicit hate and minority targets. Addressing these challenges requires better handling of culturally grounded hate and rare target categories, highlighting important directions for Persian hate-speech detection.

Limitation

First, although the dataset spans ten years, it is restricted to Twitter and may not generalize to other Persian-speaking platforms such as Instagram or Telegram, where discourse norms and interaction patterns differ. Second, the dataset exhibits substantial class imbalance, particularly the dominance of political hate. While this reflects naturally occurring discourse, it constrains model performance on minority target categories and limits balanced evaluation across identity dimensions. Moreover, target classification performance remains considerably lower than binary hate detection, especially for low-frequency identity groups. This highlights the structural difficulty of multi-target modeling under severe imbalance and implicit reference. A more comprehensive investigation of target-specific modeling strategies, balancing techniques, and architectural adaptations lies beyond the scope of the present work and remains an important direction for future research. Third, implicit hate constitutes a relatively small proportion of the dataset (7%), which restricts fine-grained modeling and detailed statistical analysis of implicit rhetorical strategies. Although we provide error analysis for both hate and multi-label target prediction, a deeper qualitative investigation would further illuminate model failure patterns. Finally, all experiments focus on supervised evaluation within ParsHate and cross-temporal transfer from PHATE. Broader cross-domain, cross-platform, and cross-cultural generalization remain to be explored. Future work may expand implicit hate coverage, incorporate additional platforms, and conduct more extensive qualitative analyses of implicit rhetorical strategies.

References

Hyeseon Ahn, Youngwook Kim, Jungin Kim, and Yo-Sub Han. 2024. Sharedcon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10444–10455.

Taha Shangipour Ataei, Kamyar Darvishi, Soroush Javdan, Amin Pourdabiri, Behrouz Minaei-Bidgoli, and Mohammad Taher Pilehvar. 2022. Pars-off: a benchmark for offensive language detection on farsi social media. *IEEE Transactions on Affective Computing*, 14(4):2787–2795.

Zewen Bai et al. 2025. STATE ToxiCN: A benchmark for span-level target-aware toxicity extraction in Chinese hate speech detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Zahra Bokaei, Walid Magdy, and Bonnie Webber. 2025. Culture matters in toxic language detection in Persian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9290–9304, Vienna, Austria. Association for Computational Linguistics.

Agostina Calabrese, Tom Sherborne, Björn Ross, and Mirella Lapata. 2025. Compositional generalisation for explainable hate speech detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13932–13954.

Hadi Davardoust, Hadi Zare, and Hossein Rafiee Zade. 2024. The dark side of instagram: A large dataset for identifying persian harmful comments.

Thomas Davidson, Dana Warmsley, 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 512–515.

Zahra Delbari, Nafise Sadat Moosavi, and Mohammad Taher Pilehvar. 2024. Spanning the spectrum of hatred detection: A persian multi-label hate speech dataset with annotator rationales. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17889–17897.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2022. Impact and dynamics of hate and counter speech online. *EPJ data science*, 11(1):3.

Faeze Ghorbanpour, Daryna Dementieva, and Alexander Fraser. 2025. Data-efficient hate speech detection via cross-lingual nearest neighbor retrieval with limited labeled data. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29662–29680.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset

for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Danial Kamali, Behrooz Janfada, Mohammad Ebrahim Shenasa, and Behrouz Minaei-Bidgoli. 2022. Evaluating persian tokenizers. *arXiv preprint arXiv:2202.10879*.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Liu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2023. Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16235–16250. Association for Computational Linguistics.

Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. 2020. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Tony McEnery and Andrew Hardie. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. 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 41–58, Online. Association for Computational Linguistics.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Xinyue Shen, Yixin Wu, Yiting Qu, Michael Backes, Savvas Zannettou, and Yang Zhang. 2025. Hatebench: Benchmarking hate speech detectors on llm-generated content and hate campaigns.(2025).

Mohammad Karami Sheykhlan, Jana Shafi, and Saeed Kosari. 2023. Pars-hao: Hate speech and offensive language detection on persian social media using ensemble learning. *Authorea Preprints*.

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott A. Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico. Association for Computational Linguistics.

Manuel Tonneau, Diyi Liu, Niyati Malhotra, Scott A Hale, Samuel Fraiberger, Victor Orozco-Olvera, and Paul Röttger. 2025. Hateday: Insights from a global hate speech dataset representative of a day on twitter. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2297–2321.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

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 (Long and Short Papers)*, pages 1415–1420. Association for Computational Linguistics.

A Appendix

A.1 Per-Strategy Analysis of Implicit Hate

This section expands on the per-strategy implicit-hate analysis referenced in Section 4.4. LLaMA-3-FT's F1 on implicit hate decomposes by rhetorical strategy as follows: cultural/historical reference F1 = 2, sarcasm/figurative language F1 = 4, other indirect strategies F1 = 5. Gemma-3-FT exhibits the same ordering across strategies, indicating that the gradient is a robust cross-model pattern. All three values are far below the explicit-hate F1 of 84.

This gradient mirrors the type and weight of external knowledge each strategy demands. Cultural and historical references encode hostility entirely in the referent, specific events, public figures, or political symbols, with no intrinsic textual cue; recognition requires Persian-specific event and entity knowledge. Sarcasm and figurative language retain partial structural cues such as negation, exaggerated affect, or dehumanising comparison, which enable occasional detection but are systematically defeated when the irony depends on contradiction with extra-textual context. The other-indirect category is comparatively the most tractable, as it typically retains some lexical hostility marker, exclusionary phrasing, generalisation-via-quantifier, or collective accusation, that is localisable from the text alone. The bottleneck is therefore not only implicitness per se but the knowledge load that each strategy imposes.

Notably, cultural/historical reference is both the hardest strategy and the most frequent in ParsHate, accounting for 48% of implicit-hate instances (versus 36% for sarcasm/figurative and 16% for other indirect). The model fails most precisely on the strategy that dominates the implicit-hate distribution. This skew toward historical and political reference is itself a Persian-specific finding: it reflects the structure of Persian online discourse, which is heavily anchored in shared political and religious memory, and is broadly distinct from the strategy distributions reported in English implicit-hate research (ElSherief et al., 2021).

Three representative cases, ordered from text-localisable to maximally knowledge-dependent, illustrate this gradient (LLaMA-3-FT and Gemma-3-FT both predict non-hate for all three):

(a) Other indirect. "Those people aren't Iranian; they all act for the internet-filtering government." Hostility is conveyed through identity denial, excluding a group from the national identity, and collective accusation, without slur, metaphor, or named referent. The exclusionary phrasing is localisable from the surface text alone.

(b) Sarcasm/figurative. "When people like Zarif wag their tails like loyal dogs for the U.S., what is left to say?" The expression uses animalisation as a sarcastic label; in Iranian cultural context, calling someone a dog is a strong insult implying low status and servitude, constituting dehumanisation. The structural cue (the dehumanising comparison) is text-internal, but the hostile reading depends on culturally specific affective weight.

(c) Cultural/historical reference. "Now they'd better not say anything about religion or faith anymore, when they hug Chávez's mother, this whole crowd is all the same!" The hostility depends on knowing the politically contested moment when Mahmoud Ahmadinejad embraced Hugo Chávez's mother at the latter's funeral, an event that became contested in Iranian religious discourse. Without this background, the tweet reads as a neutral generalisation.

A.2 Span-Detection Analysis

Beyond tweet-level classification, ParsHate's span-level rationales enable fine-grained evaluation of whether models can *localise* hate. Following the toxic-spans paradigm of Pavlopoulos et al. (2021), with related span-level hate-speech evaluations in HateXplain (Mathew et al., 2021) and STATE ToxiCN (Bai et al., 2025), we evaluate span identification on all gold-hate tweets using token-level F1, with character-level annotations converted via Hazm tokenisation consistent with Section 4.2.

On LLaMA-3-FT, overall span F1 = 44, decomposing into F1 = 58 on explicit hate and F1 = 14 on implicit hate. Gemma-3-FT shows the same pattern. The 58 → 14 gap reflects a qualitative shift in failure type, not merely a quantitative drop. On explicit hate, errors are predominantly *boundary errors*: the model identifies the hate region but selects a span that is too narrow, too broad, or fragmented relative to the gold annotation. Partial overlap still earns partial F1 credit, which is why explicit F1 lands at 58 rather than near zero. On

implicit hate, errors are predominantly *localisation errors*: the model either fails to produce a span at all or marks a non-anchor portion of the text. These produce zero or near-zero overlap, dragging F1 down to 14.

Decomposing the failure modes for misclassified implicit-hate span predictions on LLaMA-3-FT: 80% are no-span (the model declines to localise, mirroring the 83% non-hate prediction rate on implicit hate reported earlier in Section 4.4), 12% mark the wrong anchor (typically a named entity or emotionally charged surface token rather than the rhetorical device that carries the hostility), 5% mark the whole tweet (an over-generation default), and 3% are other partial-overlap cases. The dominance of the no-span mode confirms that span-detection failure is downstream of detection failure: if the model does not recognise hate, it has nothing to localise.

Two cases illustrate the contrast between failure modes (gold spans and model predictions translated for presentation):

(i) Explicit hate, boundary error. Tweet: "Each time they leave a family in mourning, drag a family behind ICU doors, behind prison walls, into courtroom corridors, into darkness and grief, killing people, women and men alike… Damn it, you sons of dogs, where did you come from to be this animalistic?" The gold annotation covers two non-contiguous spans: (a) the full enumeration of hate actions, and (b) the closing slur-plus-dehumanisation clause. LLaMA-3-FT marks only three short fragments inside (a) ("leave a family in mourning", "behind prison walls", "killing people") and misses (b) entirely. This is a classic under-coverage failure where the model locates roughly where hate is present but cannot capture the full rhetorical arc, and under-counts when multiple non-contiguous hate spans co-occur.

(ii) Implicit hate, localisation error. Tweet: "But honestly, when Raisi gets emotional and speaks angrily, he really gives off a sense of killing." Hate is encoded entirely in the rhetorical phrase *"gives off a sense of killing"* (a colloquial Persian construction implying menace and dehumanisation). The gold span is exactly this clause. LLaMA-3-FT produces no span at all, it does not register the rhetorical anchor as hate because, taken literally, the phrase contains no overt hostility marker.

The span-detection failure on implicit hate provides direct evidence at sub-tweet granularity: even when the input is restricted to known hate tweets, i.e., the model is implicitly told "hostility is present, find it", localisation still fails approximately 80% of the time. The hate is not in any localisable token; it is in the relationship between the text and an external referent (an event, a public figure, or a culturally entailed implication). Therefore the annotated rationales in ParsHate are therefore not only an evaluation resource but a potential supervision signal for future rationale-aware or knowledge-augmented training.

A.3 Annotator Demographics

All annotations were carried out by three female native Persian speakers, aged 25–35, who each lived in Iran for over twenty years. All hold graduate degrees (two in mathematics, one in computer science). Annotators were compensated at a fixed rate. They were informed before the task that the dataset contains hateful and potentially distressing content, were free to take breaks at any point, and could opt out at any stage. The annotation guidelines (approximately five pages, prepared in Persian; English translation in the supplementary materials) and annotation interface will be released publicly with the dataset.

A.4 Full Disagreement Statistics

This section expands on the disagreement analysis summarised in Section 3.3.

Per-strategy frequencies and disagreement rates. Among the three implicit strategies, cultural or historical reference is the most common in ParsHate (48% of implicit-hate tweets) and the most disagreement-prone (38% of cases), followed by sarcasm or figurative language (36% frequency, 27% disagreement) and other indirect strategies (16% frequency, 22% disagreement). Per-strategy Fleiss' κ values show the inverse trend: $\kappa = 0.43$ for cultural/historical reference, 0.55 for sarcasm, and 0.58 for other indirect strategies.

Rationale-span structure. Implicit-hate spans are longer than explicit ones (31% vs. 17% of tweet characters) and predominantly clause-level: 57% multi-token clauses, 29% noun phrases, and 14% mixed or multi-clause selections. Explicit spans, by contrast, rely more on short hate lexicon. Named entities appear in 43% of implicit spans versus 13% of explicit ones, reflecting references to politicians, regimes, and historical figures.

Adjudication procedure. About 16% of implicit-hate tweets and 6% of explicit-hate tweets required adjudication beyond majority voting. Resolution was two-step: multi-label assignment was allowed when multiple rhetorical strategies plausibly applied; remaining disagreements, typically on the implicit/explicit boundary or whether figurative language qualified as hate, were discussed using the guidelines and pilot examples, with unresolved ties defaulting to the majority label.

Tweet	Gold label	Model prediction
دیرو زود داره سوخت و سوز هرگز به چشم فقط بهتون حق انتخاب و ازادیه بیان هم میدیم طناب چه رنگی دوست دارید اخوندای انگلیس سرشت؟ It may happen sooner or later, but burning and destruction are inevitable. We're even giving you the right to choose and freedom of speech — what colour rope do you prefer, you British-bred mullahs?	Hate; politics/religion; implicit hate (reference, sarcasm)	Not-hate
اره منتهی بعضی وقتا چندتا کامیون میهن با هم میرم حجم سفارش سرکوبگر زیاده 😂 Yeah, but sometimes I go with several Mihan trucks together — the volume of orders for suppressors/oppressors is huge 😂	Hate; politics; implicit hate (reference, sarcasm)	Not-hate
اره اروح بابات تون چهل و سه سال جون مردمو گرفتید خدا جون شماها بگیره Yeah, screw your father too. For forty-three years you've taken people's lives — may God take yours too.	Hate; politics; implicit target	Not-hate
حرومزاده ها معلوم نیس چ بلایی سره این دختر اوردید 😳 مهسا واقعا راحت شد از این همه تاریکیو بدبختی روحت شاد Bastards, it's unclear what you did to this girl 😳 Mahsa is finally free from all this darkness and misery. Rest in peace.	Hate; politics; implicit target	Hate; target = gender
معلومه که رنگینتر هست اون اخوند هست و مردم ادم 👉 👉 👉 Obviously, it's more colourful/brighter — he is a mullah, and the people are human. 👉 👉 👉	Hate; religion; implicit hate + implicit target (sarcasm)	Not-hate

Table 6: Annotator rationale spans on representative *all-years* misclassifications.

Type	Tweet	Outcome
Explicit (correct)	«واقعاً متأسفم، اما همهٔ آخوندها چیزی جز کثافت نیستند». "I'm really sorry, but all mullahs are nothing but filth."	Explicit hostility and identity targeting; correctly classified as hate by both LLaMA-3-FT and Gemma-3-FT.
Implicit, figurative (missed)	«زرافه‌ها حیوونای بی‌آزاری‌ان، ولی یه زرافه تو سوریه تا الان جون چهل هزار سوری رو گرفته. علوفهٔ این زرافه هم از جیب مردم ایران تأمین می‌شه». "Giraffes are harmless animals, but one giraffe in Syria has so far taken the lives of forty thousand Syrians. The feeding trough of this giraffe is funded by the Iranian public treasury.	Hate is expressed through dehumanising metaphor referencing an unnamed political or military actor in the Syrian conflict; the literal surface contains no explicit hate cue. Predicted as non-hate by both LLaMA-3-FT and Gemma-3-FT.
Implicit, reference (missed)	«من که دیگه این کلاس‌ها رو نمی‌رم، به شما هم پیشنهاد می‌کنم نرید—آخرش می‌شید یه «استاد برجسته» مثل سعید طوسی». "I'm not going to attend these classes, and I suggest you don't either—because in the end you'll turn into some 'high-ranking teacher' like Saeed Toosi."	Interpretation depends on contextual knowledge of the named figure and recognition of the sarcastic use of 'high-ranking'. Predicted as non-hate by both LLaMA-3-FT and Gemma-3-FT
Implicit, reference (model contrast)	«حالا دوباره حسنِ «کلیدساز» و رفقاش برگردن و با همون «حال خوب» کشور رو آباد کنن—همین‌طور کلیدش رو هی بالا و پایین کنه». "Let Hassan the 'key-maker' and his buddies come back again and bring some 'good vibes' to our country—waving his key up and down all the time."	Mockery depends on recognising the 'key' as former President Hassan Rouhani's campaign symbol. Correctly flagged as hate by GPT-5 but missed by LLaMA-3-FT and Gemma-3-FT.
Implicit, figurative (GPT-5 false Positive)	«برجام اون‌قدر توافق فوق‌العاده‌ای بود که هنوز داریم دنبال نتایجش می‌گردیم». "Barjam was such a great deal that we're still searching for its results."	Gold-labelled as non-hate (critical but non-hostile commentary about the JCPOA nuclear deal, locally known as Barjam). LLaMA-3-FT and Gemma-3-FT correctly predict non-hate, while GPT-5 over-flags hostility.

Table 7: Representative cases illustrating the explicit–implicit performance gap

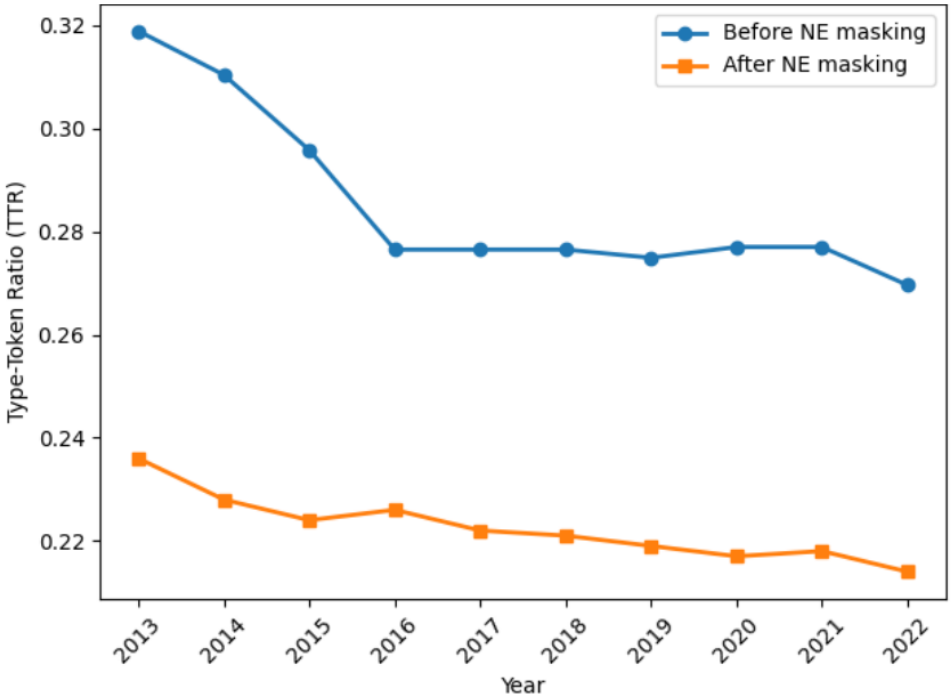

Figure 4: Lexical variability (TTR) before and after named-entity masking across years.

Hate Tweet	Model	Gold label	Predicted Label
یادتونه دریافت مدال توسط ناخدای ناو وینسنت رو که ؟؟ خب حالا یادتون باشه ما هم این پرواز رو نه میبخشیم نه فرتموش میکنیم!! . | Do you remember when the captain of the USS *Vincennes* received a medal? Well, remember this too — we will neither forgive nor forget this flight.	All year-Llama 3	Hate – implicit - Reference	Non-hate
[تحریم، کرونا، ... پرواز ۷۵۲، کشته های کرمان، آبان ۹۸ ...] فردا قشنگ است " به قشنگی همین روز هامویی که برامون ساختن!!! | [Sanctions, COVID, Flight 752, the victims of Kerman, November 2019...] Tomorrow is beautiful — as beautiful as these days they've created for us!!!	GPT 5	Hate – implicit - Sarcasm	Non-hate
تنفری که من از این قشر دارم از آخوند و سپاهی ندارم. در واقع دشمن اصلی ما " کسانی هستند که درد و فقر و شر را عادی سازی میکنند." | The hatred I feel toward this group is even greater than what I feel toward the clerics and the IRGC. In fact, our real enemies are those who normalize pain, poverty, and evil.	All year-Llama 3	Hate - 'Other' as target	Hate and 'Politics'
شما کاسه‌لیسان ولایت هم بزودی باید پاسخگو باشید | You sycophants of the Supreme Leader will soon be held accountable as well.	GPT 5	Hate – 'Politics' as target	Hate and Religious

Table 8: Misclassification samples for the best models.

A.5 Annotation Guidelines

The full annotation guidelines provided to annotators are reproduced below in English translation; the original Persian version will be released publicly with the dataset.The aim of these guidelines is to provide a structured procedure for identifying and categorising hate speech in Persian tweets. They define clear criteria for distinguishing hate speech from violent or merely offensive content, explain how to identify the targeted individual or group, and describe how each target is assigned to a predefined category. They also distinguish between explicit and implicit hate speech and provide key cues for recognising indirect forms of hate, such as figurative language or references to historical and cultural events or symbols. Annotators are expected to apply these guidelines carefully and consistently.

Step 1 – Identifying the Presence of Hate Speech. A tweet should be labelled as containing hate speech if it expresses or encourages hatred toward an individual or group on the basis of their affiliation with a specific category such as political views, occupation, religion, nationality, ethnicity, gender, age, disability, or similar.

Distinguishing hate speech from violence and offensive language. It is important to distinguish hate speech from violent or offensive content. While such content can be harmful, it does not always fall within the definition of hate speech. In other words, a tweet may contain violent or offensive content without being hateful.

Violent content includes the following:

- Threats of violent acts against a specific individual or group.
- Expressing wishes, hopes, or desires for the death or serious physical harm of others, or inciting such outcomes.
- Inviting or encouraging others to harass an individual or group.

Example. "We will make your living and dead bodies tremble; this time you won't escape death, you dug your own grave." This tweet contains violence (a threat of violent action against a specific group) but does not contain hate speech.

Offensive language includes the use of vulgar or profane language without targeting a group based on identity, or insulting a specific individual without framing them as a representative of a broader group.

Example. "This Mahdi Tootonchi guy is putting on an act, or he really is just that idiotic." The tweet uses only vulgar and offensive language; there is no hate speech.

Note that hate speech can overlap with violent or offensive language, but not every violent or offensive expression constitutes hate speech.

Example. "Bastards, you executed people to scare us? Nothing matters to me anymore, I'll stain my hands with your filthy and impure blood." This contains *hate speech* (for describing the hateful act of execution), is also labelled as *offensive* (for the vulgar term "bastards"), and contains *violence* (for "I'll stain my hands with your blood").

Step 2 – Explicit vs. Implicit Hate Speech. If a tweet conveys hate indirectly, using non-literal language, it is classified as *implicit hate speech*. Such language typically avoids overt hate vocabulary. When a tweet is labelled as implicit hate, the *strategy* by which the hate is conveyed must also be specified (figurative language, historical/cultural reference, or other indirect means). Three strategies are considered for implicit hate:

(i) Figurative language: the use of symbolic and non-literal expressions to convey hate, such as irony, sarcasm, or exaggerated comparisons.

Example. "These people are stuck to the country's resources like leeches again." A group of people is figuratively compared to leeches, bloodsuckers, implicitly conveying exploitation, parasitism, or harm to society. The hate is carried without any explicit hate vocabulary.

Example. "Hassan the 'key-maker' has been so clever and precise in making his keys that the elites are fleeing the country one by one! Well done!" On the surface, the sentence describes Hassan as a clever and meticulous key-maker, but the tone is clearly mocking. The speaker sarcastically attributes the emigration of elites to his performance, indirectly invoking his incompetence. Sarcasm allows the speaker to convey hostility or anger without using overt hate vocabulary; this makes the sentence a form of implicit hate speech in which the negative message is delivered through irony.

(ii) Reference to historical or cultural events or symbols: the use of events, figures, or symbols associated with hatred to convey a hateful message

without expressing it directly.

Example. "We don't want them to repeat November 2019 for our young people again, right?" Without naming any individual or group, the speaker refers to a historical event (November 2019) that, in public memory, is associated with repression, protests, or state violence. The reference implicitly conveys warning, threat, or distrust, all of which contribute to hate, toward a political group or governing body. Such historical or cultural references can indirectly evoke hate or political/social opposition and produce a polarising atmosphere, even when the surface of the sentence appears supportive or restrained.

(iii) Other: any other indirect means of conveying hate that does not fall under figurative language or historical/cultural reference.

Note. A tweet labelled as implicit hate may use more than one strategy simultaneously, in which case multiple options may be selected. For example, the "Hassan the key-maker" tweet above uses both figurative language and historical/cultural reference (alluding to former President Hassan Rouhani's display of a key during one of his interviews).

Step 3 – Identifying the Target of Hate Speech. If a tweet is identified as hate speech, the next step is to determine the targeted individual or group. The target can fall into one of the following predefined categories:

Gender. Hate directed at individuals or groups on the basis of their gender or gender identity.

Example. "It was these same women who, by going around without their proper clothing, turned everyone else into trash! Get lost, and now they also want the right to divorce and whatnot!"

Note. If gender is the target, specify whether the target is *women* or *men*.

Politics. Hate directed at political ideologies, parties, politicians, or political supporters.

Example. "Folks, do you hear our children's voices from Evin? Do you realise that if this regime lasts another year, we'll essentially no longer hear the name Iran?"

Note. If politics is identified as the target, specify whether the target is *domestic* or *foreign*. The tweet above targets the Iranian regime (domestic). In contrast, "Monarchists residing in England have no right to demand! For years they have been making our situation worse and fishing in muddy waters!" targets monarchists residing in England (foreign).

Religion. Hate directed at individuals or groups because of their religious beliefs or affiliations.

Example. "So statues are forbidden in the clergy's interpretation of Islam, but when it comes to their own criminals it suddenly becomes permissible, they truly do worship false idols!"

Note. If religion is the target, specify whether the target is *Islam* or *another religion*. In the example above the target is Islam.

Nationality. Hate expressed on the basis of country of origin or citizenship.

Example. "Do you get it that Afghans have ruined Iran?"

Note. If nationality is the target, specify which nationality is targeted (Iranian, Afghan, Arab nations, Western nations, or other).

Occupation. Hate that targets people because of their profession or occupational role.

Example. "I've never seen a more rotten group than doctors in Iran! Freeloaders and rude! Total scum!"

Ethnicity/Race. Hate based on race, ethnicity, or perceived family background.

Example. "Even with private tutoring, when a Lor keeps failing, honestly, education isn't their fault, they really don't get it!"

Other. If the target of hate does not fit any of the categories above, the *Other* option is selected.

Note. A tweet may contain multiple targets; in that case all applicable options are selected.

Step 4 – Explicit vs. Implicit Targets. As in Step 2, annotators indicate whether the target of hate is mentioned directly or invoked indirectly through non-literal language.

Example. "The tortured and pain-stricken bodies, the spilled blood at the hands of #Zahhak, are the price paid for freedom, no room for grief!!! Until we are struck down and martyrs are made, there is no freedom. We have no time to mourn, this is the time for war!" Instead of directly naming a specific person, the word *Zahhak* (a mythological tyrant in Persian literature) is used.

Example. "These bastards I hate, they made it to parliament again!" A pronoun (*these bastards*) is used instead of directly naming the hated individuals.

Step 5 – Identifying the Hate Span. Annotators mark all the most minimum yet specific parts of the tweet where the hate is expressed using square brackets [ ].

Example. “In every language, declare [the IRGC a terrorist organisation] and demand [the expulsion of the regime’s ambassadors].”

Example. “These same religious people who, the moment something is forbidden and illegal, oppose it and [tear apart whoever does that thing], the moment that thing becomes permissible for any reason, they themselves are first in line to do it. They [don’t oppose anything on rational grounds]; their only justification is its being forbidden [a clear example is that they were the first to send their wives to the corrupt and morally compromised sports stadium environment].”

Recording the Annotations. All responses are recorded as binary values (0 or 1), where 1 indicates presence or an affirmative response and 0 indicates absence or a negative response.

Summary of the Labelling Procedure.

1. Does this tweet contain hate speech? *If not, stop.*
2. Is the hate expressed implicitly or explicitly (directly or indirectly)?
3. If implicit, specify the strategy (figurative language, historical/cultural reference, or other).
4. Specify the target (gender, politics, religion, occupation, nationality, ethnicity, or other), with the relevant subcategory (e.g., women/men; domestic/foreign; Islam/other; Iranian/Afghan/Arab/Western/other).
5. Is the target expressed implicitly?
6. Mark the hate span(s).