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, categorizes implicit expression strategies, 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 Paersian 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 on the challenging nature of the dataset and its diverse sampling of hate speech in Persian language, which motivates for the development of 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 rapid growth of user-generated content (Bokaei et al., 2025). Over the past decade, numerous annotated datasets have been developed to support automatic hate speech detection, particularly in English and other high-resource languages. These datasets typically provide labeled examples of hateful and non-hateful content, and sometimes include additional annotations, such as hate targets or span-level rationales, that highlight the textual evidence supporting the label (Davidson et al., 2017; Mathew et al., 2021; Kennedy et al., 2022). Some datasets also distinguish between explicit and implicit hate expressions, reflecting the increasing complexity of modeling harmful language online (ElSherief et al., 2021).

In such datasets, annotation commonly involves labeling whether a text expresses hostility and identifying the target group toward which this hostility is directed. Recent datasets have also introduced span-level rationales and more structured target taxonomies, enabling analysis of both the linguistic realization of hate and the identity groups involved (Mathew et al., 2021; Kennedy et al., 2022; ElSherief et al., 2021; Davidson et al., 2017). These annotation schemes provide richer supervision and support more detailed evaluation of hate speech detection models.

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 but relies partially on keyword-based candidate selection and covers a relatively 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 studies highlighting the importance of target-aware annotations for modeling implicit hatred (Bokaei et al., 2025), the dataset distinguishes between explicit and 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 expression and support the development of more robust target detection 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 led to the development of numerous annotated datasets across languages and platforms. In this section, we review (i) benchmark datasets developed for English and (ii) existing Persian hate speech resources. This overview situates ParsHate within the broader dataset landscape.

2.1 Hate Speech Datasets in English

Several benchmark datasets in English and other high-resource languages have shaped the development of hate speech detection systems. Regarding English, several hate speech datasets have shaped the development of automatic detection systems. Davidson et al. (2017) introduced a widely used three-class Twitter dataset distinguishing hate speech, offensive language, and neither, demonstrating that lexical approaches often conflate profanity with hate and overestimate true hate prevalence. Mathew et al. (2021) proposed HateXplain, a benchmark annotated with class labels, target categories, and token-level rationales, enabling explainable modeling and bias analysis in neural systems. (Toraman et al., 2022) constructed large-scale English and Turkish datasets balanced across five target categories (religion, gender, racism, politics, sports), showing strong Transformer performance and cross-domain transferability.

Recent work has further addressed implicit and coded hate, where hostility is conveyed without overt slurs or profanity. ElSherief et al. (2021) introduced a theoretically grounded taxonomy of implicit hate speech with fine-grained categories and target group annotations, highlighting the complexity of modeling indirect hostility. Hartvigsen et al. (2022) proposed TOXIGEN, a large-scale machine-generated dataset balanced across identity groups and designed to improve robustness against subtle and implicit toxicity. However, while these benchmarks advance modeling of implicit hate expression, they do not systematically distinguish whether the targeted group is explicitly named or only indirectly referenced in the text. In addition, (Kennedy et al., 2022) introduced the Gab Hate Corpus, a

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

27K-post dataset collected from the Gab platform, including identity-based target categories and an additional label for implicitness. In addition to the datasets discussed above, several other widely used English hate speech benchmarks have shaped the field, including HatebaseTwitter (Davidson et al., 2017; Vidgen et al., 2021; Salminen et al., 2018) which collectively provide large-scale annotations, target-aware labels, or diagnostic evaluation suites for hate detection.

2.2 Persian Hate Speech Resources

Persian hate speech resources remain comparatively limited in scale and structural depth. Pars-OFF (Ataei et al., 2022) introduced one of the earliest publicly available Persian datasets, annotated hierarchically for abusive content (hate and offensive) and target type. However, it does not model implicit hate strategies or fine-grained target categories. PHATE (Delbari et al., 2024) advanced Persian hate detection by introducing multi-label annotations for hate-related categories and providing span-level rationales. While this represents an important step toward explainability, its target annotations are provided in free-text form without a normalized taxonomy. Filter-based data collection, along with limited temporal coverage, constitutes another limitation of this dataset. Pars-HAO (Sheykhlan et al., 2023), which is not publicly available, constructed a three-class Persian dataset (hate, offensive, neutral) using keyword- and profile-based sampling and reported competitive baseline results. However, it does not provide structured target schemas, span-level rationales, or implicit labeling. PHICAD substantially (Davardoust et al., 2024) increases dataset scale and provides multi-label annotations for harmful content in Persian Instagram comments. However, it does not introduce a structured target taxonomy or fine-grained target categories.

Overall, existing Persian datasets provide important foundations but lack unified hate target taxonomies, modeling of implicit hate strategies, and longitudinal coverage, thereby limiting fine-grained identity-aware and temporal analysis. Prior work on Persian hate speech detection has also highlighted the difficulty of modeling implicit hatred (Bokaei et al., 2025; Delbari et al., 2024). For this reason, datasets that explicitly annotate implicit hate expressions are important for providing appropriate supervision for training and evaluating detection models. In addition, prior studies in high-resource languages (Mathew et al., 2020; Garland et al., 2022) suggest that patterns of hateful discourse can vary over time, motivating the need for datasets that enable longitudinal analysis. However, existing Persian resources cover relatively limited temporal spans, making it difficult to examine how hate expression and identity targets evolve in Persian online discourse. These limitations highlight the need for a Persian dataset that provides structured target annotations, models implicit hate expressions, and supports analysis across longer temporal spans.

Table 1 summarizes how our ParsHate dataset fills in the limitations of the existing datasets.

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 first extracted all Persian tweets from Twitter archive, which spans between 2013 and 2022 have a sample 1% of the Twitter stream from every day. The archive contained approximately 47 million Persian tweets posted between 2013 and 2022. After removing retweets, duplicate posts, tweets shorter than five tokens, and tweets containing only URLs, mentions, or hashtags, the corpus was reduced to 29 million tweets. This filtered pool served as the source for

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 ملچک!! 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 شد! خودش و طرفداراش مملکت رو به خاک سیاه کشوندن! 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

Figure 1: Dataset instances with corresponding labels.

the hybrid temporal sampling procedure described below.

To enable systematic sampling across different 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: Based on the definition proposed by (Waseem and Hovy, 2016), we define hate as content that expresses or encourages hostility toward individuals or groups based on their association with identity-related categories. This includes discrimination, dehumanization, negative stereotyping, humiliation, or encouragement of harm directed at social or political groups. Each tweet was first labeled at the tweet level for hate (binary: hate / 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. Gender distinguishes between men and women. Religion distinguishes Islam and other religious identities, reflecting the empirical predominance of Islam-targeted hate in Persian discourse (Delbari et al., 2024). Politics distinguishes between domestic and foreign political actors. Nationality requires annotators to specify the targeted national group when identifiable. When target of hate is implied without being explicitly mentioned, it gets annotated as "implicit target".

Implicit Hate: Following prior work distinguishing 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 (implicit target), distinguishing between cases where the attacked group is directly named versus indirectly invoked (Kennedy et al., 2022)

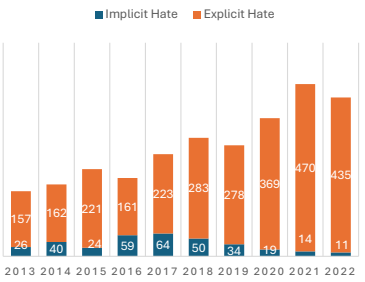

Figure 2: Hate Samples Distribution Across years.

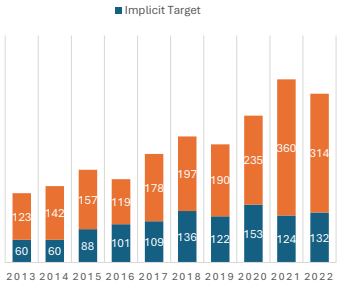

Figure 3: Implicit-Explicit Hate Target Distribution Across years.

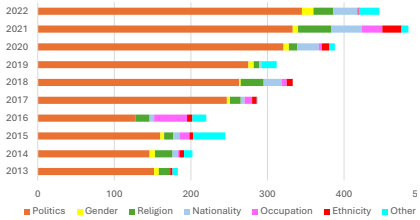

Figure 4: Hate Target Distribution Across Years.

Span-Level Annotation: For tweets labeled as conveying hate, annotators highlighted the shortest segment(s) of text that justify the assigned labels. 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 tweets were independently annotated by three native Persian speakers following detailed written guidelines (reported in the Appendix). A pilot phase was conducted to manage interpretations and refine definitions, particularly for implicit hate and implicit targeting. Inter-annotator agreement was measured using Fleiss' κ. Agreement scores were $\kappa = 0.69$ for hate identification, $\kappa = 0.60$ for target category,$\kappa = 0.51$ for implicit hate, and $\kappa = 0.58$ for implicit target. Final labels were assigned using majority voting, with discussion-based resolution in cases of persistent disagreement.

For span annotations, agreement was computed at the character level, yielding $\kappa = 0.55$. Final instance-level labels were assigned using majority voting, with discussion-based resolution in cases of persistent disagreement.

3.3 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 2, 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 3 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 4 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) target detection behavior under class imbalance. Together, these analyses provide insight into the challenges posed by the dataset for hate and target detection 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. Following (Bokaei et al., 2025), all our tuning experiments used LLama 3. 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; (ii) integrated training on 80% of data from all ParsHate years ("ALL-Year"), followed by evaluation on each year separately; and (iii) year-specific training, where models are trained and tested within the same year.

Table 2 presents the performance of all experiments. The integrated ALL-Year 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.

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

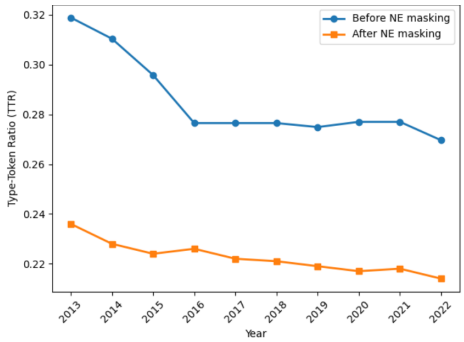

Figure 5: Lexical variability (TTR) before and after named-entity masking across years.

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 5 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 Target Detection

We next examine target detection performance. As shown in Figure 4, political targets account for 76.45% of hate instances, while other categories (e.g., occupation, ethnicity, gender) 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 target detection analysis on them. We further assess robustness by

Year	Models																	
	SOTA			GPT-5			ParsBERT			LLaMA 3			Gemma 3			ALL-Year		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
2013	**65**	51	57	54	**66**	59	X	X	X	42	40	41	43	45	44	64	63	**64**
2014	**68**	49	57	51	62	56	X	X	X	48	42	45	55	38	45	67	**66**	**67**
2015	69	61	65	50	50	50	X	X	X	63	59	61	60	56	58	**79**	**75**	**77**
2016	**72**	74	73	66	62	64	X	X	X	60	51	55	54	56	55	71	**79**	**75**
2017	73	65	69	58	56	57	61	52	56	70	62	66	60	68	64	**81**	**79**	**80**
2018	75	77	76	68	72	70	64	62	63	70	68	69	74	72	73	**78**	**80**	**79**
2019	82	**88**	85	76	74	75	59	63	61	62	64	63	66	64	65	**90**	86	**88**
2020	78	**84**	81	**88**	76	**82**	65	69	67	67	71	69	69	73	71	77	81	79
2021	**93**	**88**	**91**	67	71	77	78	79	77	85	77	81	82	84	83	85	87	86
2022	91	85	88	**93**	87	**90**	79	79	79	76	82	79	79	75	77	89	**91**	**90**
Average	77	72	75	67	68	68	68	69	68	64	62	63	64	63	63	**78**	**79**	**79**

Table 2: Precision (P), Recall (R), and F1 (F) scores across years and models

Target	Models											
	LLaMA-3-FT			GPT-5			LLaMA (Mask)			GPT-5 (Mask)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Gender	18	13	15	**22**	**80**	**34**	11	7	8	21	**80**	33
Politics	84	96	**89**	96	32	48	82	**98**	**89**	**97**	27	43
Religion	65	32	**43**	14	**81**	23	**71**	12	21	14	**81**	24
Nationality	43	8	13	7	**60**	12	**57**	10	**17**	8	**60**	14
Occupation	20	2	4	5	60	9	**33**	3	**33**	8	**65**	14
Ethnicity	**15**	2	4	4	60	8	14	2	6	6	**65**	**11**
Other	**46**	6	11	30	22	26	33	2	4	25	**33**	**28**
Macro avg	42	23	27	25	56	23	**43**	19	**27**	26	**59**	24

Table 3: Per-target performance comparison across models with and without masking.

comparing performance before and after NE masking.

As represented in Table 3 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 (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 3 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 detection settings.

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 فراموش 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 روزهاموپی 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

Figure 6: Misclassification samples for the best models.

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-Year model (LLaMA-3-FT) and compare its behavior with GPT-5 as the next strong model. We focus separately on hate detection and target detection. Table 6 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-Year 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.

Target Detection Error: Target detection 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-Year 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.

Together, these findings show that implicitness is a central driver of modeling difficulty in ParsHate. Even high-performing models struggle to reliably resolve identity categories when hostility is encoded indirectly or when targets lack sufficient training representation. Taken together, these analyses show that ParsHate presents both structural and temporal challenges for hate and target detection in Persian. Temporal evaluation reveals a clear distribution shift, with models trained on recent discourse struggling to generalize to earlier years characterized by higher lexical variability. Lexical analysis further shows that early years exhibit greater diversity, only partly explained by named entity frequency, suggesting broader topical variation over time. Structurally, severe target imbalance—particularly the dominance of political hate—creates asymmetric learning conditions that bias models toward majority-class predictions. Error analysis further indicates that implicit hate expressions and implicit targets account for many misclassifications. Overall, these results show that temporal variation, lexical diversity, class imbalance, and implicitness jointly shape the core modeling challenges in ParsHate.

4.5 Moving Forward

The results on ParsHate show that temporal variation, implicit hate, and strong target imbalance remain challenging for current models. Addressing these issues will require approaches that better handle indirect and culturally grounded expressions of hate, as well as improved modeling of low-frequency target categories. Future work should also consider how to build models that remain stable across time and how to evaluate them under such conditions. We hope that making *ParsHate* publicly available encourages further work in these directions and contributes to more reliable hate speech detection in Persian.

5 Conclusion

We introduce ParsHate, the first decade-spanning (2013–2022) Persian benchmark for hate and target detection. Comprising 10,000 manually annotated tweets, the dataset provides a structured multi-label target taxonomy, explicit–implicit distinctions, strategy categories, and span-level rationales. By employing score-stratified temporal sampling, ParsHate preserves natural label distributions while mitigating keyword-driven bias. Our evaluation reveals significant temporal and structural challenges; specifically, models trained on contemporary discourse struggle to generalize to earlier years characterized by higher lexical variability. While multi-year training improves robustness, it does not fully eliminate performance gaps in early-year data. Furthermore, error analysis indicates that implicit strategies and indirect, culturally grounded target references remain primary drivers of misclassification. These findings underscore the persistent difficulty of modeling identity-directed hate when expressed through evolving or indirect linguistic structures, and motivates future research in this direction.

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 target detection, 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

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.

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.

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.

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.

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.

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.

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*.

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.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

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.

A Appendix