Spanning the Spectrum of Hatred Detection: A Persian Multi-Label Hate Speech Dataset with Annotator Rationales

Zahra Delbari[1], **Nafise Sadat Moosavi**[2], **Mohammad Taher Pilehvar**[3]
[1] Tehran Institute for Advanced Studies, Khatam University, Iran
[2] University of Sheffield, United Kingdom
[3] Cardiff University, United Kingdom
z.delbari@khatam.ac.ir, n.s.moosavi@sheffield.ac.uk, pilehvarmt@cardiff.ac.uk

Abstract

With the alarming rise of hate speech in online communities, the demand for effective NLP models to identify instances of offensive language has reached a critical point. However, the development of such models heavily relies on the availability of annotated datasets, which are scarce, particularly for less-studied languages. To bridge this gap for the Persian language, we present a novel dataset specifically tailored to multi-label hate speech detection. Our dataset, called **PHATE**, consists of an extensive collection of over seven thousand manually-annotated Persian tweets, offering a rich resource for training and evaluating hate speech detection models on this language. Notably, each annotation in our dataset specifies the targeted group of hate speech and includes a span of the tweet which elucidates the rationale behind the assigned label. The incorporation of these information expands the potential applications of our dataset, facilitating the detection of targeted online harm or allowing the benchmark to serve research on interpretability of hate speech detection models. The dataset, annotation guideline, and all associated codes are accessible at `https://github.com/Zahra-D/Phate`.

1 Introduction

Hate speech is a pervasive problem in our societies which is particularly amplified by social media.

Hate speech encompasses various forms of communication that specifically target individuals or groups, with the intention of attacking, dehumanizing, inciting violence, promoting discrimination, or fostering hostility. These harmful actions are based on perceived characteristics such as race, ethnicity, religion, gender, sexual orientation, political beliefs, social class, disability, or appearance. Expressions of this nature seek to demean, marginalize, or inflict harm on others, creating a hostile environment and perpetuating hatred, prejudice, and inequality. The lasting impact on mental health at the individual level, coupled with the societal-level consequence of normalizing prejudice and discrimination, underscores the critical need for developing effective methods to identify and address hate speech.

In response to this pressing need, a substantial body of research has emerged focusing on the automated detection of hate speech. Researchers have delved into various facets of hate speech, encompassing aggression (Kumar et al. 2018), cyberbullying (Chen and Li 2020), harassment (Stoop et al. 2019), racism (Field et al. 2021) and sexism (Chiril et al. 2020).[1] However, many of these solutions heavily rely on the availability of annotated datasets for training or evaluation, a resource that is often lacking, particularly for low-resource languages such as Persian.

To our knowledge, there are only three datasets for hate speech detection for this language (Ataei et al. 2022; Mozafari, Farahbakhsh, and Crespi 2022; Alavi, Nikvand, and Shamsfard 2021). Unfortunately, among these, only ParsOFF (Ataei et al. 2022) is accessible to the public. While the dataset remains a valuable resource, it is essential to acknowledge its reliance on a weak heuristic during the data collection stage, resulting in a relatively trivial set of instances. Notably, over half of the tweets labeled as offensive in this dataset were collected using a simple search for offensive keywords. This methodology introduces potential bias towards these specific terms, thereby achieving a high accuracy in distinguishing offensive from non-offensive posts, even with a straightforward BERT-based model.

In this paper, we aim to bridge this gap by introducing a Persian hate speech dataset that is curated based on a more rigorous procedure. Referred to as PHATE, this dataset comprises over 7,000 Persian Twitter posts, each manually annotated by two annotators. The dataset follows a hierarchical structure, with posts initially classified at the top level as either normal or hate. The hate category is further divided by at least one of the three possible sub-labels: violence, hate, and vulgar. Consequently, the task in PHATE is characterized as a multi-level multi-label classification setup.

Beyond the hate speech tags, PHATE extends its scope by incorporating information about the target of the hateful conduct, which may pertain to an individual or a group. This additional feature not only enriches the dataset but also opens avenues for various downstream applications. For example, it enables initiatives to address online violence directed at specific groups, such as women journalists. Moreover, each instance in PHATE is accompanied by human ra-

[1] While there have been some attempts to unify the taxonomy and set of labels in different datasets (Fortuna, Soler, and Wanner 2020), hate speech remains a highly subjective task, leading to significant variations across existing work.

tionales explaining the corresponding label assignment. This feature renders the dataset as a valuable resource for research in interpretability, facilitating a deeper understanding of models' intuitions behind specific labels in the context of hate speech classification. This can also serve as a resource for aligning the thought processes of models with human rationales.

We evaluated various models on PHATE to set baseline. The observed 10-20% performance gap of the best monolingual Persian pre-trained (ParsBERT) and multilingual (XLM-R) models underscores the challenging nature of the benchmark, providing impetus for future research in Persian hate speech detection. It is noteworthy that ChatGPT (zero-shot) consistently performs below these baselines across all categories. In addition, we conducted different experiments based on the provided rationales, aiming to underscore their reliability for research on interpretability and their potential to assist models in their decision-making processes.

Along with PHATE, we provide transparent label definitions and annotation guidelines, which can pave the way for the construction of similar datasets in other languages or for the integration of the dataset into broader multi-lingual datasets.

2 Related Work

Given that our dataset intersects various research domains, we address the related work for each of these in distinct subsections.

2.1 Hate Speech Subcategories

The prevalent approach to frame hate speech detection is a three-way classification: categorizing the input content as hate speech, offensive language, or neither (Toraman, Şahinuç, and Yilmaz 2022; Mulki et al. 2019; Luu, Nguyen, and Nguyen 2021; Mathew et al. 2021; Davidson et al. 2017). However, the definitions of the Hate, Offensive, and Normal categories are not consistently aligned across these studies. Additionally, there are researches that encompasses other facets of hate, including aggression (Kumar et al. 2018), cyberbullying (Chen and Li 2020; Sprugnoli et al. 2018; Cheng et al. 2021; Nitta et al. 2013), harassment (Stoop et al. 2019; Ghosh Chowdhury et al. 2019), racism (Field et al. 2021; Hasanuzzaman, Dias, and Way 2017; Davidson, Bhattacharya, and Weber 2019), and sexism (Chiril et al. 2020; Sen et al. 2022; Costa-jussa et al. 2021). Additionally, Zampieri et al. (2019) introduced a hierarchical structure for labeling, encompassing categorizing data as hate or normal. For data classified as hate, this approach also considers whether it's targeted and if so, what type of target it belongs to, whether it's an individual group or others.

2.2 English Hate Speech Datasets

A wide range of datasets are available for hate speech detection that are primarily focused on the English language. These datasets are derived from various sources including Twitter (ElSherief et al. 2021; Bianchi et al. 2022; Davidson, Bhattacharya, and Weber 2019; Davidson et al. 2017; Waseem and Hovy 2016; Founta et al. 2018), Reddit (Kurrek, Saleem, and Ruths 2020; Vidgen et al. 2021), YouTube (Sarkar and KhudaBukhsh 2021; Hammer et al. 2019), Instagram (Zhong et al. 2016; Suryawanshi et al. 2020), and Facebook (Luu, Nguyen, and Nguyen 2021; Kumar et al. 2018). Among these sources, Twitter used to be one of the most popular due to its convenient API for academic research, which is no longer available.

2.3 Hate Speech Datasets in Other Languages

Hate speech datasets are not limited to English. Due to the significance of this task, there is a growing number of datasets with hate speech annotations available in various languages, including Korean (Yang, Jang, and Cho 2022; Lee et al. 2022), Arabic (Alsafari, Sadaoui, and Mouhoub 2020; Hadj Ameur and Aliane 2021), German (Wiegand, Siegel, and Ruppenhofer 2018; Struß et al. 2019), Indonesian (Ibrohim and Budi 2019), Vietnamese (Luu, Nguyen, and Nguyen 2021), and Bengali (Romim et al. 2021). However, the number of datasets available for these languages is relatively smaller compared to those annotated for English.

Persian is one of the low-resource languages in this regard. The existing datasets for Persian hate speech detection include Pars-OFF (Ataei et al. 2022), and two other non-public datasets introduced by Mozafari, Farahbakhsh, and Crespi (2022), and Alavi, Nikvand, and Shamsfard (2021). Pars-OFF comprises 7,381 normal and 3,182 offensive Persian tweets, organized into a three-level hierarchy as outlined in Zampieri et al. (2019).The process of collecting tweets employed a combination of similarity-based and keyword-based data selection strategies. Mozafari, Farahbakhsh, and Crespi (2022) curated a dataset of 6k Persian tweets adhering to the annotation guidelines established in Zampieri et al. (2019). The tweets were collected through two distinct methods: random sampling and lexicon-based sampling, over a two-months period spanning from June to August 2020 from Iranian platforms such as Digikala and Snappfood. Alavi, Nikvand, and Shamsfard (2021) created a dataset named POLID, which includes 2,453 entities labeled as "NOT" and 2,535 as "OFF". Instances were labeled "OFF" if they contained certain offensive words, with subsequent adjustments made to account for potential misclassifications due to polysemous words or implicit offensive language.

2.4 Multilingual Hate Speech Datasets

Although many datasets focus on a single language, there are also multilingual datasets available. For example, the dataset introduced by Ousidhoum et al. (2019) covers Arabic, French, and English, providing annotations for hate speech detection across these languages. Additionally, Toraman, Şahinuç, and Yilmaz (2022) present a large dataset that encompasses both English and Turkish. Another notable dataset is the SemEval dataset introduced by Basile et al. (2019), which includes hate speech annotations for both Spanish and English. None of these datasets cover Persian.

2.5 Annotation of Hate Speech Rationales

While hate speech classification has been extensively researched, there are limited studies that specifically provide and analyze the spans corresponding to hate labels. Annotating the rationales behind hate speech in datasets has several benefits for hate speech detection research. It enables the analysis of linguistic patterns and contextual cues indicative of hate speech that can lead to developing more accurate detection models (Mathew et al. 2021). Additionally, it facilitates the development of interpretable models, enhancing transparency and accountability by identifying the rationales for each hate speech label.

Notable examples of datasets that annotate hate speech rationale spans include ViHOS (Hoang et al. 2023), the SemEval-2021 Task 5 dataset (Pavlopoulos et al. 2021), DOSA (Ravikiran and Annamalai 2021), and HateXplain (Mathew et al. 2021). ViHOS introduces comprehensive guidelines for identifying hate-related spans in Vietnamese comments. It contains 30,000 YouTube and Facebook comments, labeled as normal, hate, or offensive. The SemEval-2021 Task 5 dataset provides toxic spans for 10,629 posts derived from the Civil Comments dataset (Borkan et al. 2019). DOSA presents a dataset of offensive spans in low-resource languages, comprising 4786 Tamil-English and 1097 Kannada-English YouTube comments, each annotated with offensive segments. The HateXplain dataset focuses on hate and offensive spans at the word level. It contains 20,148 Gab and Twitter posts, with each post manually classified into hateful, offensive, or normal categories.

3 Dataset Construction

Twitter, a widely used social media platform, serves as a platform for millions of users worldwide to share their thoughts and opinions on various topics. However, the open and transparent nature of Twitter also fosters an environment where hate speech and online harassment can thrive, particularly targeting individuals with protected attributes or those expressing their views. In this section, we outline our approach to constructing a hate speech dataset based on data retrieved from Twitter, taking advantage of the platform's API to collect a substantial amount of relevant data.

3.1 Data Collection

Selecting candidate tweets. The selection of candidate textual content for hate speech detection datasets plays an important role in ensuring the quality and representativeness of the annotations. The chosen approach for data selection can introduce biases or limitations to the datasets. Common approaches include searching for lists of slurs and derogatory keywords (Waseem and Hovy 2016; Kurrek, Saleem, and Ruths 2020), focusing on specific events or contexts (Grimminger and Klinger 2021), or adopting a mixture of strategies (Basile et al. 2019; Ataei et al. 2022; Fersini, Nozza, and Rosso 2018).

To minimize biases and avoid superficial artifacts in our dataset, we avoid explicitly searching for slurs. Instead, we curate a list of (1) controversial words that elicit mixed opinions among the public, such as the names of well-known celebrities with divergent public perceptions, and (2) ambiguous words that have different meanings in different contexts, e.g., the word رژیم can either refer to "diet" or "regime" (as in government), with the latter interpretation being a potential target for hateful comments.

We used the Full-Archive Search feature provided by Twitter's Premium API to search for tweets containing our specific keywords, primarily focusing on the period from 2020 to early 2023. We specifically selected Persian tweets that were original posts, excluding retweets and replies, and did not contain any media attachments. Through this process, we obtained a total of 4,179,797 tweets, which we then narrowed down to 3,882,984 by removing duplicates.

Data filtering. To ensure a balanced representation of hate speech and normal tweets in our final dataset, we employed a systematic approach. We uniformly sampled 810 tweets for each keyword and employed ChatGPT[2] (Brown et al. 2020) for the initial classification into hate speech or normal categories.[3] Despite our intention to categorize the data into two classes, for some input, ChatGPT may not offer a distinguishable answer. Hence, alongside 'normal' and 'hate,' we introduced a third category, 'cannot identify.' We utilized this 'cannot identify' category to avoid potential biases that might be introduced by ChatGPT. To achieve this, we equally selected data from all three classes. Additionally, we ensured an equal selection of tweets for each keyword. Using this approach, we selected approximately 7,000 tweets for further annotation by annotators.

3.2 Annotation Guidelines

To ensure reliable and accurate annotations, it is essential to define clear labels and guidelines for annotators. This is particularly important given the dynamic nature of hate speech where its definition remains ambiguous. Moreover, having clear definitions for each label will allow other datasets to easily map their labels, despite any discrepancies or inconsistencies in label names. This mapping process can help ensure a common understanding of what each label means and reduce redundancy in future work (Fortuna, Soler, and Wanner 2020).

In our case, as we primarily use tweets, we consulted the Twitter Rules and policies[4] to begin defining our labels. This guide outlines Twitter's stance against two main categories: any form of violent behavior ranging from threatening comments to wishes of harm, and any behavior that could be construed as hateful towards a specific set of protected groups. These concepts are widely used in other hate speech datasets, so we decided to adopt them as subcategories for our dataset as well. However, since a tweet can simultaneously contain both types of hate and violence, a multi-label scheme was necessary.

We also break down hate speech directed at a group or an individual (not representing a group) into a separate subcategory, based on Jeremy Waldron's (Waldron 2012) argument

[2] https://chat.openai.com/

[3] The classification was performed using the specific prompt *Does the following text contain hate speech? [tweet text].*

[4] https://help.twitter.com/en/rules-and-policies/violent-entities

Tweet Text (Pr): از کلینیک زیبایی باهام تماس گرفتن که مدل عکس قبل از عمل شم
Tweet Text (En): *They contacted me from the beauty clinic to ask me to be the model for the 'before' photo of the surgery.*
Label(s): Normal

Tweet Text (Pr): آقای ترامپ قمارباز حداقل تو این دوماه باقیمونده کامران نجف زاده رو اعدام کن تو کشورت
Tweet Text (En): *Mr. Trump, the gambler , at the very least, during these two months that are left, execute Kamran Najafzadeh in your country.*
Label(s): Violence , Vulgar
Traget(s): Trump, Najafzadeh

Tweet Text (Pr): بابا مسلمانی یه فحشه، کی میخوایین اینو بفهمید. سوال درست اینه گورخر وحشی تانزانیایی بهتر حکومت میکنه یا یک مسلمان؟
Tweet Text (En): *Being a Muslim is a curse , when are you going to understand this? The real question is, does a Tanzanian wild zebra rule better or a Muslim?*
Label(s): Hate
Target(s): Muslims

Figure 1: Examples of dataset instances with corresponding labels.

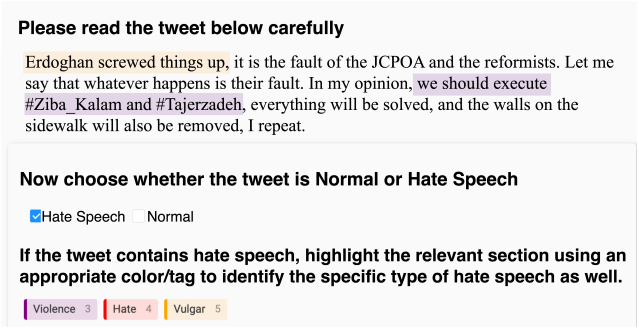

Figure 2: An example of span annotation (translated to English)

that this type of hate speech can be more harmful than hate speech directed at an individual, as it can create a climate of intolerance and prejudice that affects the entire group and harm innocent members.

Therefore, we considered the following three categories (Figure 1 presents an example for each of the labels):

- **Violence** includes (1) threats of violent acts against an identifiable target; (2) wishing, hoping, promoting, inciting, or expressing a desire for death, serious physical harm; and (3) calling for and encouraging others to harm or harass.
- **Hate** includes hate or its encouragement, directed at individuals or groups based on their association with specific categories such as politics, occupation, religion, or race. Often, this targeting focuses on distinguishing attribute of the group, such as age, gender, race, nationality, religion, or disability. Hate can include, but is not limited to, discrimination, negative stereotype, dehumanization, and humiliation.
- **Vulgar** includes profanity speech without any specific target or towards an individual who is not a representative of a general group.

A tweet is considered as hate speech if at least one of the above three categories applies.

3.3 Annotation Procedure

Pilot Test. We conducted a pilot test on 300 instances to evaluate the consistency and clarity of the guidelines as well as to identify reliable annotators. As a result, parts of the initial guidelines were refined to reduce potential points of confusion. We also selected two annotators (both graduate students) who had high consistency in their annotations and demonstrated alignment with the intended objectives of the guidelines.

Annotation instructions. The selected annotators were instructed to perform the following tasks:

1. Determine whether the tweet is normal or contains any form of hate speech.
2. If the tweet contains hate speech, determine all the hate speech categories, based on the provided definitions.
3. Highlight the exact part of the text that supports the labels you chose in step 2 (using their corresponding colors).
4. If the target of the hate is known, please specify it.

We employed the Label Studio[5] platform for annotating the data. The annotation environment is depicted in Figure 2. Prior to the annotation process, we informed the annotators about the possibility of encountering profanity language in the data. To ensure privacy and anonymity, we replaced user IDs in tweets with the placeholder "USER".

Target and rationale. Besides labeling the tweets using the provided labels, we also requested the annotators to denote the target of the hate speech and the shortest spans of the tweets that implied the selected hate speech label. The target are specified in a free text form. In addition, the annotators were asked to indicate the identity dimension of the target group from a set of pre-defined options: Political, Racial, Religious, Occupational, National, and Gender labels. For example, if the target group in a tweet labeled as hate speech is "Muslims", the corresponding identity dimension would be selected as "Religious Group". Regarding data labeled as vulgar, the target can either be an individual or there may be no specific target at all. The former is sometimes referred to as cyberbullying in other research studies,

[5]https://app.heartex.com/

Label	F-1 Agreement	Kappa
Hate Speech	0.88	0.72
Violence	0.60	0.58
Hate	0.68	0.59
Vulgar	0.68	0.61

Table 1: Agreement statistics for the two annotators.

Number of Tweets	7,056
Number of Normal Tweets	3,860
Number of Hate Labels	1,632
Number of Vulgar Labels	1,583
Number of Violence Labels	582
Number of Tweets Contain Violence and Hate	241
Number of Tweets Contain Hate and Vulgar	236
Number of Tweets Contain Violence and Vulgar	148
Number of Tweets Contain All Three Labels	23

Table 2: Label distribution statistics in the PHATE dataset.

while the latter is categorized as an offensive language without a specific target.

In order to gather data for potential usage of the dataset in research on interpretability, we also instructed the annotators to highlight spans that they considered as the rationale behind their decision. Section In 4.4, we utilize this rationale to improve the performance of speech detection models. Figure 2 represents an annotated example with its corresponding rationale in the annotation platform.

3.4 Quality Assessment

To ensure accurate labeling, we implemented a two-step process. In the first step, both annotators independently labeled the tweets. If there was any disagreement between them in at least one label, a third annotator, who is one of the authors, acted as an adjudicator. The adjudicator reviewed the tweets where there was disagreement and provided the final label based on their judgment.

Gold annotations. Regarding the gold label, it is possible that two annotators may choose different targets and spans. In such cases, we report both annotators' choices for further usage and analysis. It is important to note that the tweets adjudicated by the third annotator, have only one annotation in terms of span and target, ensuring consistency in the annotation process.

Hate Speech	Violence	Hate	Vulgar
Stupid 140	Kill 71	Dirty 107	*Mortazavi* 98
Dirty 120	Fire 71	Regime 92	*Rahimpoor* 93
Mortazavi 117	Regime 54	Stupid 78	*Dabir* 88
Regime 115	Die 42	Muslim 61	*Torabi* 82
Rahimpoor 105	Dirty 24	Officer 59	*Totonchi* 76

Table 3: The list of the five most frequent initial keywords for each label category, with the corresponding tweet counts. Keywords in italics are the names of individuals.

Inter-annotator agreement. We calculated the inter-annotator agreement between the two annotators for the binary classification task of determining whether a tweet is normal or hate speech using the Fleiss' Kappa metric. The resulting Fleiss' Kappa score was 0.72, indicating a substantial level of agreement. The Fleiss' Kappa score for sub-labels are presented in Table 1. Furthermore, we computed the agreement accuracy for the hate speech label. If both annotators agreed on labeling a tweet as hate speech, it was considered a hit, and vice versa. The accuracy was 0.85, demonstrating the clarity of the guidelines and the reliability of the manual annotations.

Human performance. In order to estimate an upperbound performance for PHATE, we randomly selected 500 ($\sim 0.07\%$) tweets. To ensure an unbiased evaluation, we involved a fourth person who had not previously participated in the annotation process. The annotator was asked to tag the selected tweets based on the defined labels. By incorporating a new annotator, we aimed to assess the reliability of our dataset beyond the initial annotations. The results of this evaluation are discussed in Section 4.2.

3.5 Dataset Statistics

The statistics regarding the dataset and labels are presented in Table 2. It is important to note that the violence, hate, and vulgar labels are not mutually exclusive, which means there are tweets that contain two or even all three of these labels simultaneously.

Table 3 displays the contribution of the five most frequent initial keywords for each label. It is worth noting that in the case of vulgar tweets, which can have an individual target or no target at all, the most frequent keywords consist of the names of popular celebrities and well-known individuals.

4 Experiments and Results

The following four models were used in our experiments.

ParsBERT. Developed by Farahani et al. (2021), ParsBERT is a Persian monolingual language model based on the BERT architecture (Devlin et al. 2019). Trained on over 73 million Persian sentences from diverse sources, it employs the same objective function as the original BERT model and is served as the primary pretrained LM for Persian.

mBERT. It is the multilingual variation of the BERT model (Devlin et al. 2019) that is pretrained on over 100 languages, including Persian.

XML-R. Conneau et al. (2020) presented XML-R, a multilingual model based on the RoBERTa architecture. Trained on a large dataset of 2.5TB filtered CommonCrawl text, which includes Persian content, this model utilizes an objective function akin to RoBERTa's.

ChatGPT. This renowned model is based on the GPT architecture developed by OpenAI (Brown et al. 2020). Specifically, for our experiments, we utilize the OpenAI API with the gpt-3.5-turbo model.

Model	Hate Speech			Violence			Hate			Vulgar		
	Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1
ParsBERT	80.5±2	75.9±1	78.1±1	42.8±3	**68.2**±5	52.3±1	59.4±2	**62.5**±1	60.8±1	**68.4**±8	**55.0**±4	**60.3**±1
mBERT	78.5±3	72.5±2	75.3±1	41.7±4	62.8±5	49.7±1	64.6±6	55.1±3	59.0±1	67.3±4	48.5±1	56.3±1
XML-R	**80.8**±6	**76.4**±5	**78.1**±1	**50.0**±4	62.8±7	**55.1**±1	**66.8**±7	58.1±5	**61.6**±1	63.0±5	54.7±4	58.2±1
ChatGPT	55.2	77.6	64.3	85.5	23.0	36.2	85.0	32.5	47.0	50.3	44.7	47.3
Human	95.3	79.7	**86.8**	74.0	84.9	**78.7**	51.3	52.3	51.8	83.7	63.5	**72.2**

Table 4: Performance of different models in detecting hate speech as well as its three specific sub-labels. The last row indicates the performance of human evaluator.

4.1 Experimental Setup

We split the dataset into the train, validation, and test sets with the respective proportions of 50%, 10%, and 40%. Since the dataset is highly imbalanced in terms of the Violence, Hate, and Vulgar labels, we use the F1 score of the test set to compare the models. Considering the multi-label nature of these sub-labels, separate evaluations were performed for each, resulting in four binary tasks when combined with the hate speech vs normal classification.

We utilize ChatGPT as a zero-shot learner. However, since our label names are general words, we provide the definition of each label along with the tweet text to ensure a more accurate evaluation. We then prompt ChatGPT to determine whether the tweet contains the corresponding label.[6] If ChatGPT's response lacked a definite answer, we iterated the prompt until achieving a classification response that aligns with the desired outcome.

For the other three models, during the training phase, they were trained on the training set utilizing the AdamW optimizer, with a learning rate of 2×10^{-5} and a batch size of 32. These models were trained for 10 epochs, and the reported test dataset results are based on the epoch that achieved the highest F1 score on the validation set. To ensure result robustness, we repeated the experiment with five random seeds (except for the ChatGPT model, which does not require fine-tuning) and reported their mean values, as shown in Table 4. The human performance results were obtained from the evaluation explained in Section 3.4.

4.2 Baseline Performance

The experimental outcomes are detailed in Table 5, revealing a substantial performance gap between state-of-the-art models and human proficiency. These findings underscore the inherent intricacy of the task, even considering that human performance in the hate category is not exceptionally high. By meticulously examining evaluator annotations, we can infer that the comparatively lower precision in the Vulgar category arises from instances where the tweet mentions a group using a vulgar adjective. In these cases, the evaluator often categorizes them as employing profanity language (Vulgar label) rather than indicating hate toward that group (Hate label). Consequently, the precision of the Vulgar category diminishes, and since these instances should indeed be marked as hate, there is a reduction in the recall of the Hate category as well. Regarding the low precision of the Hate category, it can be said that evaluator had a stricter threshold to classify criticism as hate, as opposed to annotators.

Among the diverse models assessed, ParsBERT and XML-R showcase well-rounded performance across various labels, with XML-R outperforming ParsBERT in all labels except Vulgar by a subtle margin. ParsBERT, being a Persian monolingual model, exhibits good performance possibly due to its specialization in the Persian language. On the other hand, XML-R demonstrates notable multilingual capabilities, which contribute to its strong performance.

Among the fine-tuned models, mBERT displays the weakest performance despite sharing architecture and parameter count with the other two models. ChatGPT lags behind the other three models, indicating minimal system bias introduced by ChatGPT during initial data filtering. The model's low precision across all labels could be attributed to its zero-shot learning approach. While we provided definitions for each sub-label, its reliance on general terms might lead to categorizing more content as violence, hate, or vulgar based on its own conceptual understanding of these words.

4.3 Evaluating Rationales

To verify the validity of the annotated rationales, a validation process was undertaken. For each sub-label, tokens corresponding to the rationales of that specific sub-label, i,e, Violence, Hate, Vulgar were masked within the input sentences. Subsequently, a ParsBERT model that fine-tuned for the sub-label was utilized to compute the F1 score for this rational-masked test set. For comparison purposes, an alternate version of this test set was generated, where an equivalent number of words as in the corresponding rationale were randomly chosen to be masked. It is evident to mention that in both test sets, the inputs with normal label are not masked. The model's performance on this test set was also assessed. The outcomes of this experimental procedure are detailed in Table 5. The results therein illustrate a notable performance decrease for the rationale-masked test set in contrast to the randomly-masked counterpart. This discrepancy underscores the significance of the annotated rationales in influencing model performance. In the absence of these ratio-

[6]The exact prompt was *Considering the definition of [sub-label] as "[sub-label definition in Subsection3.2]" does the following text contain [sub-label]? "[text]" please answer in the classification format. True for yes and False for no.*

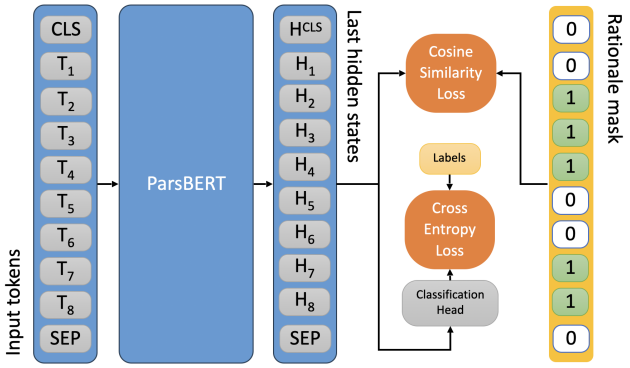

Figure 3: Rationale-assisted fine-tuning process.

nales, which we define as the segments of sentences contributing to hate expression, the model's predictions of hate speech categories would be substantially impacted. Particularly noteworthy is the substantial drop in recall, as we anticipate, given that the model shouldn't be able to predict a sentence with rationale masking as hate speech.

4.4 Rationale-assisted Fine-tuning

Utilizing the rationales, we enhance the performance of the ParsBERT model, reaffirming the effectiveness of annotated rationales. Motivated by the work of Mathew et al. (2021) and Camburu et al. (2018), where they employed explanations pertinent to the task to enhance performance, our goal is to employ rationales for strengthening our model's understanding of the task.

Based on the provided rationales, we create masks for the tokens within the input sentences. Afterwards, each batch goes through two folds of training. In the first fold, we calculate an additional loss to encourage the representation of the last hidden layer of the BERT model to become more similar to the mask. This is achieved by computing the cosine similarity between the rationale mask and the norm one of BERT output hidden states. This additional loss is calculated only for instances that pertain to the specific sub-label for which the model is being trained, and the underlying BERT model is updated using this loss. In the second fold, we update the entire model (BERT + classification head) in a conventional manner using the classification loss. The entire training process is illustrated in Figure 3. The results of this training are displayed in Table 6, indicating performance improvement.

5 Conclusions

In response to the scarcity of resources for hate speech detection in Persian, we introduce PHATE, a new dataset tailored for this purpose. Comprising 7,000 annotated tweets, the dataset includes hate speech-related labels, rationales, and targets for hateful comments. High inter-annotator agreements signify clear guidelines and consensus among annotators. A carefully devised strategy for selecting candidate posts has resulted in a relatively balanced dataset for

		Rec.	Prec.	F1
Violence (24%)	Rationale	6.75±2	24.7±2	10.4±2
	Random	36.3±3	64.6±5	46.2±1
Hate (27%)	Rationale	19.9±2	35.7±2	25.5±2
	Random	53.3±2	59.9±2	56.4±1
Vulgar (27%)	Rationale	30.9±8	34.9±2	32.0±5
	Random	62.1±8	52.6±4	56.3±1

Table 5: ParsBERT performance on randomly-masked and rationale-masked test sets. Results are reported for five experiment repetitions using different seeds. Numbers in parentheses show the mean percentage of rationale length.

		Rec.	Prec.	F1
Violence	FT	42.8±3	68.2±6	52.3±2
	FT + Rationale	**46.5**±5	**68.3**±5	**55.0**±3
Hate	FT	59.4±2	62.5±2	60.8±1
	FT + Rationale	**66.5**±5	**60.9**±3	**63.3**±1
Vulgar	FT	**68.4**±8	55.0±5	**60.3**±1
	FT + Rationale	63.7±4	**57.5**±2	60.3±1

Table 6: The results of rationale-assisted fine-tuning of ParsBERT that are reported for five different random seeds.

the classification task, distinguishing hate speech from normal discourse. We demonstrate that integrating rationales into the model could improve performance, suggesting their utility as guiding factors for the model's attention. Notably, the inclusion of hate spans within the dataset holds promise for enhancing model interpretability in future research. With a vision to contribute to the field of hate speech detection in Persian, we believe PHATE stands as a valuable resource for advancements in model development and understanding within this linguistic context.

Limitations

PHATE was collected from a specific social media platform and may not be representative of hate speech on other platforms or in other contexts. Additionally, while all the construction steps of our dataset are transferable to other languages, the dataset and results are exclusively confined to the Persian language, as they are intricately tied to Persian keywords and cultural nuances. Also, our annotation process relied on human annotators, which may introduce subjective biases and inconsistencies in the labeling. We attempted to mitigate this limitation by using multiple annotators and providing clear annotation guidelines, but some degree of subjectivity may still exist. Our approach to hate speech detection may not capture all forms of hate speech, as the definition of hate speech can vary across cultures and communities. Finally, our dataset focuses on detecting hate speech, and it cannot serve as a resource for exploring the underlying causes or contextual factors that contribute to the production of hate speech (Adewumi et al. 2022).

References

Adewumi, T.; Vadoodi, R.; Tripathy, A.; Nikolaido, K.; Liwicki, F.; and Liwicki, M. 2022. Potential Idiomatic Expression (PIE)-English: Corpus for Classes of Idioms. In *Proceedings of the 13th LREC*, 689–696.

Alavi, P.; Nikvand, P.; and Shamsfard, M. 2021. Offensive Language Detection with BERT-based models, By Customizing Attention Probabilities. arXiv:2110.05133.

Alsafari, S.; Sadaoui, S.; and Mouhoub, M. 2020. Hate and offensive speech detection on Arabic social media. *Online Social Networks and Media*, 19: 100096.

Ataei, T. S.; Darvishi, K.; Javdan, S.; Pourdabiri, A.; Minaei-Bidgoli, B.; and Pilehvar, M. T. 2022. Pars-OFF: A Benchmark for Offensive Language Detection on Farsi Social Media. *IEEE Transactions on Affective Computing*.

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F. M.; Rosso, P.; and Sanguinetti, M. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th SemEval*, 54–63.

Bianchi, F.; HIlls, S.; Rossini, P.; Hovy, D.; Tromble, R.; and Tintarev, N. 2022. "It's Not Just Hate": A Multi-Dimensional Perspective on Detecting Harmful Speech Online. In *Proceedings of the 2022 EMNLP*, 8093–8099.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 WWW*, 491–500.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33.

Camburu, O.-M.; Rocktäschel, T.; Lukasiewicz, T.; and Blunsom, P. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31.

Chen, H.-Y.; and Li, C.-T. 2020. HENIN: Learning Heterogeneous Neural Interaction Networks for Explainable Cyberbullying Detection on Social Media. In *Proceedings of the 2020 Conference on EMNLP*, 2543–2552.

Cheng, L.; Mosallanezhad, A.; Silva, Y.; Hall, D.; and Liu, H. 2021. Mitigating Bias in Session-based Cyberbullying Detection: A Non-Compromising Approach. In *Proceedings of the 59th ACL and the 11th IJCNLP*, 2158–2168.

Chiril, P.; Moriceau, V.; Benamara, F.; Mari, A.; Origgi, G.; and Coulomb-Gully, M. 2020. An Annotated Corpus for Sexism Detection in French Tweets. In *Proceedings of the 12th LREC*, 1397–1403.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th ACL*, 8440–8451.

Costa-jussa, M.; Gonen, H.; Hardmeier, C.; and Webster, K., eds. 2021. *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*.

Davidson, T.; Bhattacharya, D.; and Weber, I. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the 3rd ALW*, 25–35.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1): 512–515.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 NAACL: Human Language Technologies*, 4171–4186.

ElSherief, M.; Ziems, C.; Muchlinski, D.; Anupindi, V.; Seybolt, J.; De Choudhury, M.; and Yang, D. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 EMNLP*, 345–363.

Farahani, M.; Gharachorloo, M.; Farahani, M.; and Manthouri, M. 2021. ParsBERT: Transformer-based Model for Persian Language Understanding. *Neural Processing Letters*, 53(6): 3831–3847.

Fersini, E.; Nozza, D.; and Rosso, P. 2018. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In *EVALITA@CLiC-it*.

Field, A.; Blodgett, S. L.; Waseem, Z.; and Tsvetkov, Y. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the ACL*, 1905–1925.

Fortuna, P.; Soler, J.; and Wanner, L. 2020. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the 12th LREC*, 6786–6794.

Founta, A.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Ghosh Chowdhury, A.; Sawhney, R.; Mathur, P.; Mahata, D.; and Ratn Shah, R. 2019. Speak up, Fight Back! Detection of Social Media Disclosures of Sexual Harassment. In *Proceedings of the 2019 NAACL: Student Research Workshop*, 136–146.

Grimminger, L.; and Klinger, R. 2021. Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection. In *Proceedings of the 11th WASSA*, 171–180.

Hadj Ameur, M. S.; and Aliane, H. 2021. AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News Hate Speech Detection Dataset. *Procedia Computer Science*, 189: 232–241. AI in Computational Linguistics.

Hammer, H. L.; Riegler, M. A.; Øvrelid, L.; and Velldal, E. 2019. THREAT: A Large Annotated Corpus for Detection of Violent Threats. In *2019 CBMI*, 1–5.

Hasanuzzaman, M.; Dias, G.; and Way, A. 2017. Demographic Word Embeddings for Racism Detection on Twitter. In *Proceedings of the 8th IJCNLP*, 926–936.

Hoang, P. G.; Luu, C. D.; Tran, K. Q.; Nguyen, K. V.; and Nguyen, N. L.-T. 2023. ViHOS: Hate Speech Spans Detection for Vietnamese. In *Proceedings of the 17th EACL*, 652–669.

Ibrohim, M. O.; and Budi, I. 2019. Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter. In *Proceedings of the 3rd ALW*, 46–57.

Kumar, R.; Ojha, A. K.; Malmasi, S.; and Zampieri, M. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the 1st TRAC*, 1–11.

Kurrek, J.; Saleem, H. M.; and Ruths, D. 2020. Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage. In *Proceedings of the 4th WOAH*, 138–149.

Lee, J.; Lim, T.; Lee, H.; Jo, B.; Kim, Y.; Yoon, H.; and Han, S. C. 2022. K-MHaS: A Multi-label Hate Speech Detection Dataset in Korean Online News Comment. In *Proceedings of the 29th COLING*, 3530–3538.

Luu, S. T.; Nguyen, K. V.; and Nguyen, N. L.-T. 2021. A Large-Scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts. In Fujita, H.; Selamat, A.; Lin, J. C.-W.; and Ali, M., eds., *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, 415–426.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 14867–14875.

Mozafari, M.; Farahbakhsh, R.; and Crespi, N. 2022. Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta Learning. *IEEE Access*, 10: 14880–14896.

Mulki, H.; Haddad, H.; Bechikh Ali, C.; and Alshabani, H. 2019. L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language. In *Proceedings of the 3rd ALW*, 111–118.

Nitta, T.; Masui, F.; Ptaszynski, M.; Kimura, Y.; Rzepka, R.; and Araki, K. 2013. Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization. In *Proceedings of the 6th IJCNLP*, 579–586.

Ousidhoum, N.; Lin, Z.; Zhang, H.; Song, Y.; and Yeung, D.-Y. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of the 2019 EMNLP-IJCNLP*, 4675–4684.

Pavlopoulos, J.; Sorensen, J.; Laugier, L.; and Androutsopoulos, I. 2021. SemEval-2021 Task 5: Toxic Spans Detection. In *Proceedings of the 15th SemEval*, 59–69. Online.

Ravikiran, M.; and Annamalai, S. 2021. DOSA: Dravidian Code-Mixed Offensive Span Identification Dataset. In *Proceedings of the 1st DravidianLangTech*, 10–17.

Romim, N.; Ahmed, M.; Talukder, H.; and Saiful Islam, M. 2021. Hate Speech Detection in the Bengali Language: A Dataset and Its Baseline Evaluation. In Uddin, M. S.; and Bansal, J. C., eds., *Proceedings of IJCACI*, 457–468.

Sarkar, R.; and KhudaBukhsh, A. R. 2021. Are Chess Discussions Racist? An Adversarial Hate Speech Data Set (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(18): 15881–15882.

Sen, I.; Samory, M.; Wagner, C.; and Augenstein, I. 2022. Counterfactually Augmented Data and Unintended Bias: The Case of Sexism and Hate Speech Detection. In *Proceedings of the 2022 NAACL: HLT*, 4716–4726.

Sprugnoli, R.; Menini, S.; Tonelli, S.; Oncini, F.; and Piras, E. 2018. Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying. In *Proceedings of the 2nd ALW*, 51–59.

Stoop, W.; Kunneman, F.; van den Bosch, A.; and Miller, B. 2019. Detecting harassment in real-time as conversations develop. In *Proceedings of the 3rd ALW*, 19–24.

Struß, J. M.; Siegel, M.; Ruppenhofer, J.; Wiegand, M.; and Klenner, M. 2019. Overview of GermEval Task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, 352–363.

Suryawanshi, S.; Chakravarthi, B. R.; Arcan, M.; and Buitelaar, P. 2020. Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text. In *Proceedings of the 2nd TRAC*, 32–41.

Toraman, C.; Şahinuç, F.; and Yilmaz, E. 2022. Large-Scale Hate Speech Detection with Cross-Domain Transfer. In *Proceedings of the 13th LREC*, 2215–2225.

Vidgen, B.; Nguyen, D.; Margetts, H.; Rossini, P.; and Tromble, R. 2021. Introducing CAD: the Contextual Abuse Dataset. In *Proceedings of the 2021 NAACL: HLT*, 2289–2303.

Waldron, J. 2012. *The Harm in Hate Speech*. Harvard University Press, JSTOR.

Waseem, Z.; and Hovy, D. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, 88–93.

Wiegand, M.; Siegel, M.; and Ruppenhofer, J. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *14th Conference on Natural Language Processing (KONVENS 2018)*, 1.

Yang, K.; Jang, W.; and Cho, W. I. 2022. APEACH: Attacking Pejorative Expressions with Analysis on Crowd-Generated Hate Speech Evaluation Datasets. In *Findings of the ACL: EMNLP 2022*, 7076–7086.

Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; and Kumar, R. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 NAACL: HLT*, 1415–1420.

Zhong, H.; Li, H.; Squicciarini, A.; Rajtmajer, S.; Griffin, C.; Miller, D.; and Caragea, C. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. In *Proceedings of the 25th IJCAI*, 3952–3958.