Phate Dataset

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

Zahra Delbari, Nafise Sadat Moosavi, Mohammad Taher Pilehvar

Motivation

The research addresses the urgent need for effective hate speech detection models in online communities, given the alarming rise of hate speech. This demand is particularly pressing for less-studied languages like Persian, where annotated datasets are scarce. To fill this gap, we introduce a novel dataset called **PHATE**, specifically tailored to **multi-label** hate speech detection in Persian tweets. **PHATE** consists of over seven thousand manually -annotated tweets, each specifying the targeted group of hate speech and including a rationale behind the label assignment. By incorporating this additional information, **PHATE** facilitates the detection of targeted online harm and serves as a valuable resource for research on interpretability of hate speech detection models. Evaluation of various models on **PHATE** underscores its challenging nature and highlights the need for future research in Persian hate speech detection.

Hate

Involves hate or encouragement directed at individuals or groups based on specific categories like politics, occupation, religion, and targets distinctive group attributes. Includes discrimination, negative stereotypes, dehumanization, and humiliation.

23%

Violence

8%

Threats involving violent acts against an identifiable target, expressing a desire for death or serious physical harm, and calling for or encouraging others to harm or harass.

7,056 Tweets

3,860 Normal

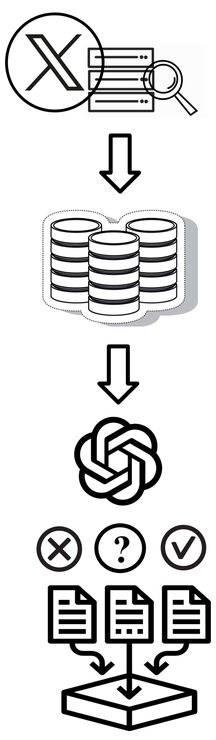
22%

Vulgar

Includes profanity speech without any specific target, and profanity speech towards an individual who is not a representative of a general group.

Dataset Construction

Data Collection



53 **controversial** query words

3,882,984 Tweets

Prompt: Does the tweet contain hate speech?

7,056 Data to annotate

Annotation Procedure

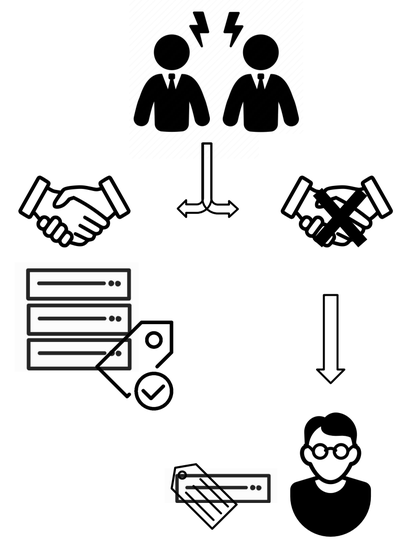
- ➔ Determine whether the tweet is normal or contains any form of hate speech.
- ➔ Determine all the hate speech sub-labels that the tweet belongs to.
- ➔ Highlight the specific portion of the text that substantiates the chosen labels.
- ➔ Specify the target of the hate speech if it is identifiable.

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

Quality Assessment



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

Experiments

➔ There is a substantial performance gap between state-of-the-art models and human performance. These findings underscore the inherent intricacy of the task, even considering that human performance in the hate category is not exceptionally high

Model	Hate Speech			Violence			Hate			Vulgar		
	Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1
ParsBERT	80.5 ± 2.9	75.9 ± 1.8	78.1 ± 0.5	42.8 ± 3.2	**68.2** ± 5.8	52.3 ± 1.8	59.4 ± 2.3	**62.5** ± 1.7	60.8 ± 0.7	**68.4** ± 8.3	**55.0** ± 4.9	**60.3** ± 0.7
mBERT	78.5 ± 3.5	72.5 ± 2.9	75.3 ± 0.4	41.7 ± 4.8	62.8 ± 5.1	49.7 ± 1.8	64.6 ± 6.9	55.1 ± 3.8	59.0 ± 1.2	67.3 ± 4.3	48.5 ± 1.8	56.3 ± 0.7
XML-R	**80.8** ± 6.0	**76.4** ± 5.2	**78.1** ± 0.5	**50.0** ± 4.5	62.8 ± 7.2	**55.1** ± 1.7	**66.8** ± 7.1	58.1 ± 5.0	**61.6** ± 1.4	63.0 ± 5.4	54.7 ± 4.1	58.2 ± 1.0
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**

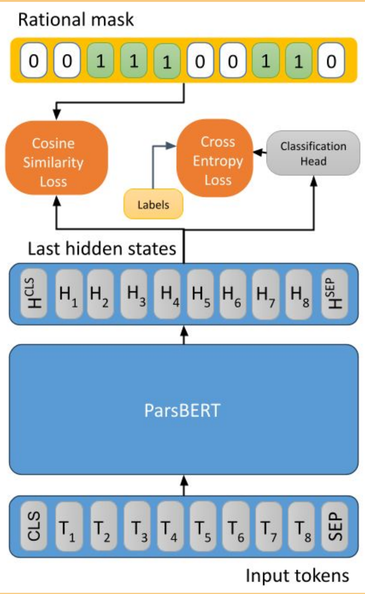
Performance of different models in detecting hate speech as well as its three specific sub-labels.

Evaluating Rationales

- ➔ ParsBERT performance on randomly-masked and rationale-masked test sets
- ➔ This discrepancy underscores the significance of the annotated rationales in influencing model performance.

		Rec.	Prec.	F1
Violence (24%)	Rationale	6.75±1.9	24.7±1.6	10.4±2.3
	Random	36.3±3.0	64.6±5.7	46.2±1.2
Hate (27%)	Rationale	19.9±2.2	35.7±2.3	25.5±2.2
	Random	53.3±1.8	59.9±1.8	56.4±0.4
Vulgar (27%)	Rationale	30.9±8.4	34.9±1.7	32.0±4.7
	Random	62.1±8.1	52.6±4.5	56.3±1.3

Rationale-assisted Fine-tuning



- ➔ Utilizing the rationales, we enhance the performance of the ParsBERT model, reaffirming the effectiveness of annotated rationales

		Rec.	Prec.	F1
Violence	FT	42.8±3.2	68.2±5.8	52.3±1.8
	FT + Rationale	**46.5**±5.1	**68.3**±5.4	**55.0**±2.7
Hate	FT	59.4±2.3	62.5±1.7	60.8±0.7
	FT + Rationale	**66.5**±5.2	**60.9**±3.4	**63.3**±0.7
Vulgar	FT	**68.4**±8.3	55.0±4.9	**60.3**±0.7
	FT + Rationale	63.7±3.6	**57.5**±2.1	60.3±0.9

Contact Info
zhr.delbari@gmail.com
https://www.linkedin.com/in/zahra-delbari-4b0356147

Read the Paper

Github Repository
https://github.com/Zahra-D/Phate