



# Persian offensive language detection

Emad Kebriyai<sup>1</sup> · Ali Homayouni<sup>1</sup> · Roghayeh Faraji<sup>1</sup> · Armita Razavi<sup>1</sup> · Azadeh Shakery<sup>1,2</sup> · Hesham Faili<sup>1,2</sup> · Yadollah Yaghoobzadeh<sup>1</sup>

Received: 14 June 2022 / Revised: 29 May 2023 / Accepted: 17 July 2023 /

Published online: 23 August 2023

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2023

## Abstract

With the proliferation of social networks and their impact on human life, one of the rising problems in this environment is the rise in verbal and written insults and hatred. As one of the significant platforms for distributing text-based content, Twitter frequently publishes its users' abusive remarks. Creating a model that requires a complete collection of offensive sentences is the initial stage in recognizing objectionable phrases. In addition, despite the abundance of resources in English and other languages, there are limited resources and studies on identifying hateful and offensive statements in Persian. In this study, we compiled a 38K-tweet dataset of Persian Hate and Offensive language using keyword-based data selection strategies. A Persian offensive lexicon and nine hatred target group lexicons were gathered through crowdsourcing for this purpose. The dataset was annotated manually so that at least two annotators investigated tweets. In addition, for the purpose of analyzing the effect of used lexicons on language model functionality, we employed two assessment criteria (FPED and pAUCED) to measure the dataset's potential bias. Then, by configuring the dataset based on the results of the bias measurement, we mitigated the effect of words' bias in tweets on language model performance. The results indicate that bias is significantly diminished, while less than a hundredth reduced the F1 score.

**Keywords** Offensive language detection · Debiasing · Imbalanced data · Twitter

## 1 Introduction

The considerable proliferation of social networking platforms and their very common use among people, along with their many applications, has created a space for hostile user accounts. As a result of the lack of effective solutions to restrict and also the presence of anonymous users, we are witnessing the spread of the phenomenon of offensive language among network users. The effects of offensive content on individuals have made researchers and the academic community determined to provide solutions to create and improve models for detecting and purging such content (Schmidt and Wiegand 2019; Founta et al. 2018; Salawu et al. 2020; Fortuna et al. 2021).

---

Editor: Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang.

Extended author information available on the last page of the article

In comparison to the extensive amount of study on hate speech detection in the English language (Waseem and Hovy, 2016; Davidson et al., 2017; Zampieri et al., 2020; Founta et al., 2018; Wiegand et al., 2019), less work has been done in other languages. People mostly interact in their native language on social media, so it is essential to provide solutions for detecting offensive content in languages other than English in order to protect their users from attack, abuse, humiliation, and insult. The availability of high-quality datasets is a crucial factor in advancing the field of offensive language detection for many languages (Fortuna and Nunes, 2018; Golbeck et al., 2017), including Persian (Alavi et al., 2021). Despite the growing interest in this area, there is still a shortage of annotated datasets that can support the development and evaluation of effective models. To address this challenge, we have developed a Persian tweet dataset for offensive language detection, which we believe is one of the largest and most comprehensive datasets in the Persian language. Our work is inspired by previous research in this area (Alavi et al., 2021; Mozafari et al., 2022), and provides a valuable resource for future studies aiming to improve the accuracy and generalizability of offensive language detection models in Persian.

Offensive language by its nature, only accounts for a small portion of total tweets and detecting such content is inherently part of the category of imbalanced learning problems. According to Founta et al. (2018), offensive tweets make up only 3% of all the tweets. The imbalanced nature of the offensive tweets data has caused the majority of datasets in previous work to be imbalanced and the models trained on them (Madukwe et al., 2020). Based on our study, approximately 3.5% of Persian tweets contain offensive and derogatory content.

Accordingly, it is not a good idea to collect tweets over a period of time and annotate them, since the number of offensive tweets would be extremely small. Therefore, inspired by Qian et al. (2019); Wulczyn et al. (2017); Waseem and Hovy (2016); Golbeck et al. (2017); Davidson et al. (2017), we gathered a collection of tweets by sampling based on keywords. We have prepared a list of offensive keywords for the Persian language. In addition to the offensive keywords, some hate keywords are collected and tweets are selected by sampling based on a combination of these keywords. The tweets are then manually labelled. This dataset contains 28K tweets divided into three categories: (1) not-offensive, (2) offensive, and (3) hate, with target groups of: religion, nationality, race, sex, etc. To the best of our knowledge, this is the first annotated dataset of tweets with offensive and hate speech labels in the Persian language.

Due to the annotation complexity and the keyword-based sampling, most of the datasets are generally biased towards specific keywords (Madukwe et al. 2020; Park et al. 2018; Dixon et al. 2018). Although many related studies train machine learning models with these datasets to detect offensive and hatespeech, the issue of system robustness due to the presence of biased data has been less studied. Dixon et al. (2018) demonstrated that ML approaches may misclassify a message because of unintended bias towards certain identity keywords.

According to our data collection method, investigations are performed for the presence of different annotation biases in the dataset. Topics that we believe the data should not be biased towards are introduced by experts, and keywords related to each topic are gathered. Then, using the FPED and pAUCED criteria proposed by Dixon et al. (2018), the bias of models' predictions relative to the identity keywords is calculated. The results demonstrate that there is a bias in the dataset towards certain identity keywords. Considering the bias-prone keywords, we proceeded to debias the dataset by selectively adding data. The results show that after this debiasing process, the mentioned criteria improve, and the dataset becomes more reliable for model use.

In recent years, various models, including ML and in particular deep learning models, are widely adopted in detection of offensive content (Mozafari et al. 2019; Badjatiya et al. 2017; Silva et al. 2016). In this paper, ML techniques such as SVM and Logistic Regression with bag of words features, as well as deep learning approaches based on pretrained language models (PLMs) are used to detect offensive language. PLMs perform well in general and the BERTweet-FA (Malekzadeh, 2020) model, trained on ~20 million Persian tweets, outperformed other models, scoring 0.903 F1 measure. This value of F1 is comparable to the best results in other languages (Zampieri et al., 2020; Fortuna et al., 2021). An experiment is also carried out to see how data imbalance affects model accuracy in terms of mild, moderate, and excessive imbalance.

The main contributions of this paper are the followings:

- We build the first manually annotated offensive tweet dataset in Persian. Moreover, we gather the largest offensive and hate speech keyword collection and use them to collect tweets.
- We investigate bias due to keyword-based sampling with two criteria and then propose selectively adding data to mitigate the bias.
- Several classic and modern NLP classification models are applied to this dataset, including different PLMs such as parsBERT and BERTweet-FA. The results show that this data is reliable for training models for Persian offensive language detection.

## 2 Related work

The flames of Hate Speech and Toxic Language are arising in the last few years due to an increasing interest in using social media, and its consequences can be led to rage, agony, and depression for the user who is reading a hateful/abusive post or comments online. So, many researchers are working on Hate Speech Detection in different languages. The literature review on this paper is categorized based on three cases: (1) Different Hate speech datasets, (2) The baseline models, and (3) The De-biasing methods and techniques applied in some research to handle the bias issue in data and model. Since our work is in Persian/Farsi hate speech and its syntax is similar to the Arabic and Urdu languages, we've covered some papers related to hate speech detection tasks in other languages and English.

*Datasets:* Many datasets for hate speech and Abusive Language are available on different platforms like Twitter, Facebook, Yahoo, YouTube, Wikipedia, Reddit, etc. For example, SemEval Contest,<sup>1</sup> which is a competition in Hate Speech Detection, collected its data from Civil Comments. Twitter is a common resource for hate speech and toxic detection. One of the largest hate speech datasets on Twitter is developed by Founta et al. (2018) and contains 100k tweets. The other large dataset is collected from ASKfm by Van Hee et al. (2015) with a size of 85k. Davidson et al. (2017) dataset, which contains 25k tweets, is another one. Kennedy et al. (2017) created a 20k multi-platform hate speech dataset, collected data from Twitter, Reddit, and The Guardian, and annotated it as harassment or not-harassment. Waseem and Hovy (2016) collected the ZeerakW Twitter dataset and labeled 17k tweets as racist, sexist, or none, and it is a good choice for working with racial hate speech detection. Wulczyn et al. (2017) developed a dataset from Wikipedia comments

---

<sup>1</sup> SemEval-2022.

containing 100k of data regarding personal attacks. HatEval dataset (Basile et al. (2019)) is a famous dataset annotated with non-hateful and hateful tweets with a size of about 13k. ETHOS is a dataset with two kinds of labeling: binary (1K comments) and multi-label (400 comments). This dataset is created from YouTube and Reddit comments by Mollas et al. (2020). There are few resources and datasets on Offensive and Hate speech detection in the Persian language; Dehghani et al. (2021) acquired one of the few available datasets containing 33k annotated tweets with Abusive and Non-Abuse labels. A 2k COVID-HATE dataset (He et al. (2021)) is another dataset related to hate speech towards Covid-19. The above-mentioned datasets are a few ones among lots of other well-known datasets which many researchers are working on for abusive and hate speech detection tasks.

*Baseline models:* Many hate speech detection pieces of research are done in English datasets. CNN+ skipped CNN, in addition to CNN+ GRU are two architectures applied to different Twitter datasets by Zhang et al. (2018). The best F1 score got by both CNN+sCNN and CNN+GRU with 94%. They also find a unique score for each tweet related to different classes. Their score defines the fraction of class-unique words in a tweet, depending on the class of that tweet. In another paper, Rajput et al. (2021) proposed a method of detecting hate speech on the Twitter dataset by using a static BERT embedding as an input of different DNNs. They concluded that working with the BERT embedding has a better performance in hate speech classification rather than other embeddings like fastText or GloVe. Their best F1 score is 79.71%, done by BiLSTM plus static BERT embedding. Speaking of BERT, Mozafari et al. (2019) fine-tuned four different BERT architectures such as BERT, BERT+Nonlinear-layers, BERT+LSTM, and BERT+CNN on different Twitter datasets, and with the fine-tuned BERT+CNN; they could get the F1 score of 92%. Chiu and Alexander (2021) did hate detection on sexism and racism tweets using the GPT3 model helping zero-shot, one-shot, and few-shot learning. With the few-shot learning on the GPT3 model, they got the Accuracy of 85% as their best result. One of the other research in this field is done by Aljero and Dimililer (2021). Working with the Stacked Ensemble approach, they used different combinations of three base classifiers, SVM, Logistic Regression, and XGBoost, with a Word2Vec feature extractor. They evaluated their models on four datasets, including HatEval and Davidson, and their best F1 score of 97% is impressive indeed. Regarding multilingual hate speech detection, Dowlagar and Mamidi (2021) used a multilingual BERT on datasets called HASOC FIRE-2020 and FIRE-2019, which contain English, German and Hindi languages from Twitter and partially from Facebook. They compared the multilingual BERT with other baseline models like SVM and ELMo+SVM, and the multilingual BERT had the best F1 score performance for all three languages. 81.5% in English, 80.4% in German, and 73.1% in Hindi subtask. We classified the Persian tweets as Offensive and Not-offensive using fine-tuning XLM BERT, Multilingual BERT, and also ParsBERT and BERTweet-FA models pre-trained on Persian Corpus and compared the results.

*De-biasing methods:* The Hate Speech Detection task is biased by its nature! Because of its complexity and being vague in what text should be considered hate or abusive, biases are inevitable. These biases are called Unintended Bias. Some papers are pointing at different bias issues on hate speech. Some people suggested the formula and metrics of Data Fairness (Hardt et al. 2016). For instance, Czarnowska et al. (2021) categorized the Fairness metrics into three generalized fairness metrics and compared different fairness metrics. Based on the Fairness definition, we can clearly understand what biases are. Followings in this section are some researchers trying to find metrics to quantify bias and come up with any methods to mitigate the bias. Some papers classify the bias based on some characteristics. For example, Shah et al. (2019) categorize the bias into four types called

Selection Bias, Label Bias, Model Over-amplification, and Semantic Bias. They also developed a predictive framework to identify where each bias might appear in the NLP pipeline. On the other hand, Garg et al. (2022) divide bias into two general categories based on the source of harm and the target of harm. Source bias includes Sampling Bias, Lexical Bias, and Annotation Bias. Target bias contains Racial, Gender, Political and other biases.

Dixon et al. (2018) introduced False Positive Equality Difference (FPED), False Negative Equality Difference (FNED), and pinned AUC Equality Difference (pAUCED) as three metrics to find the bias based on a set of terms. Kag (2019) proposed a generalized mean of the bias AUC (GMBAUC) to evaluate a set of terms exposing bias. Kennedy et al. (2020) investigated bias on imbalanced data and showed that there are some identifiers like "black" or "gay," which models are biased to them and caused the false positive. They found such biases and presented a new regularization to mitigate the biases in pre-trained BERT. Kind of the same method is done by Mozafari et al. (2020) in their previous work using BERT. First, they showed that the datasets they used had some bias, and second, using a regularization method, re-weighting the samples, and fine-tuning the BERT again could reduce racial bias in African-American English and Standard American English tweets.

Davidson et al. (2019) examined the Racial Bias in five datasets containing tweets about African American people. They showed that the model has a discriminative behavior towards these groups. Badjatiya et al. (2019) quantified bias on a set of stereotype words and used a knowledge-generalization technique for a model in order to learn the general context and mitigate the bias. Their work is done on two datasets. One was collected from Wikipedia Talk pages, and the other included Twitter data. Social Bias Frames are introduced by Sap et al. (2019) to find the social stereotypes and biases. In addition, they come up with a Social Bias Inference Corpus to support their work. Zhou (2021) proposed two kinds of de-biasing approaches. The first approach is for mitigating the known, and defined biases named the LEARNED-MIXIN method. And the other one, called Data Filtering, contains AFlite and DataMaps, for quantifying and reducing unspecified biases. Considering the lexicon-based data collection in our dataset in this paper, we applied the bias metrics FPED, FNED, and pAUCED (Dixon et al. 2018) on a set of Persian words to quantify the bias in our data. For de-biasing, after finding the biased terms, we collect the data containing that terms in the opposite class (label) to mitigate the effect of the bias on such terms.

## 2.1 Non-english hate speech detection review

Alavi et al. (2021) proposed an approach for detecting offensive content in Persian languages, which involved changing the 'Attention Mask' input and creating offensive scores based on probabilities generated by Multinomial Naive Bayes. The approach improves the performance of BERT-based models, with up to 10% improvement observed. Dehghani et al. (2021) presented the very first deep learning method using the Bert language model to detect abusive words in Persian tweets on Twitter. Their proposed method achieved an accuracy of 97.7%. Another research in the Persian Language to identify hateful content in short texts is done by Jey et al. (2022). Their proposed method uses natural language processing (NLP) techniques, including word-based and character-based n-grams and calibrated Support Vector Machine, to calculate the probability of each feature related to hate speech. Raghad Alshalan et al. (2020) collected different tweets related to Covid-19 from the ArCOV-19 dataset and labeled tweets as hate or non-hate. They also applied a topic

modeling system using an unsupervised approach to reduce the dimensionality of nonnegative matrices called NFM in order to find seven different topics of hateful tweets. Using a CNN model to classify the data, they got an F1 score of 79%. Aldjanabi et al. (2021) worked on four different Arabic Twitter datasets. They used AraBert and MarBert in Multi-task Learning. Their best result was the Multi-task learning on the MarBert model with an F1 score of 92.34% on a binary dataset labeled offensive and not offensive. Haq et al. (2020) also proposed a lexicon-based framework called USAD (Urdu Slang and Abusive words Detection) to detect abusive text in Perso-Arabic-scripted Urdu Tweets with an F1 score of about 63% and Recall of 74.3%. Their work drew our attention because, firstly, the Urdu language and Persian are similar somehow. Secondly, just like our work, due to no dataset availability, they built their abusive lexicon and the dataset and annotated it manually. One of the other works in Arabic hate speech detection on Twitter is done by Aljarah et al. (2021). They evaluated the data on four ML models: SVM, Naïve Bayes, Decision Tree, and Random Forest. Using different combinations of feature extraction techniques, they did feature importance analysis by Gini metric to extract the most important features in the data. Generally, the Random Forest showed the best result for the task compared to other models. In conclusion, we can state that the work presented in this paper, which consists of building the Persian hate speech Twitter dataset and manually annotating it, as well as our contribution to previous works, is one of the very first and most extensive to conducted in the Persian language.

### 3 Data preparation

Research in Natural Language Processing, particularly in Offensive Language and Hate Speech Detection (HSD), requires specific pre-processing libraries (normalization, sentence detection, tokenization, stemming, and lemmatization), a stop-words and offensive words dataset, and a specific definition of Offensive and Hate Speech for each language (like the specific definition of hate target groups and their related keywords). We applied pre-processing to the tweet texts in two stages: first, when sampling the data and checking against keywords, and second, before preparing feature vectors for training machine learning models (such as LR and SVM).

Persian is a low-resource language that lacks rich resources and reliable libraries in most aforementioned areas. Prior to developing a detection model for this research, we collected a broad range of Offensive and Hate Speech Detection requirements. The following is a basic summary of the process of preparing requirements and collecting datasets:

#### 3.1 Keyword collection

In this study, two types of keyword datasets are needed: Persian offensive keywords and related keywords to hate speech target groups. Collecting offensive keywords was done in two steps; the first step was using the primary offensive dataset that students of the University of Tehran collected in the 3 years for different usage and projects. The second step of expanding and completing the offensive keyword dataset is to filter the collected tweets containing these keywords, list the descending count of unigram words, remove the non-offensive words in the list as much as possible, and manually assess the remaining words to find new offensive keywords and add to the dataset if there are. The related keywords to hate speech target groups are gathered by a comprehensive search, especially institutes'

websites related to target groups and encyclopedia websites such as Wikipedia. In the next step, we did the same expanding and completing task for each group, the same as the offensive keywords dataset. The collected datasets consisted of names of villages, cities, provinces, races, tribes, and guilds of Iran and names of religions, sects, nations, countries, and related words to each group. We aggregated the collected data as follows:

- Offensive: the root of all Persian offensive words.
- Races-Places: name of all Iranian races and tribes. All villages, cities, and provinces of Iran. The names of nations, countries. Also, all related words and hashtags to these groups.
- Religions: the name of religions and sects and related words and hashtags.
- Others: the name of guilds of Iran and all remaining groups' related words and hashtags.

### 3.2 Data selection

Due to the scarcity of a dataset of Persian casual speeches, the University of Tehran regularly collects a massive collection of Persian tweets from various users. The TWINT<sup>2</sup> crawler gathered this dataset. TWINT is a Twitter crawler that enables the extraction of Tweets from Twitter without utilizing the Twitter API. As our main repository, we used a portion of the collected Persian tweets for 2020, which contained 200 gigabytes of raw tweets. We prioritized three keyword datasets: races, places, and religions, based on our field observations on Twitter and the volume of collected related terms. To procure data, three steps were taken: filtering approximately 850,000 tweets with Races-Places keywords, filtering approximately 870,000 tweets with Religions keywords, and obtaining approximately 750,000 tweets independent of context. All data selection stages were performed regardless of whether tweets overlapped; additionally, data were randomly selected from three four-month time intervals in 2020 to guarantee that data sampling includes a broad range of dates and events.

We stored selected data in a SQL-Server database and deleted tweets that were duplicated in context or by ID. Additionally, to expedite future tasks such as data labelling and modelling, we searched the whole database for tweets containing phrases from each keyword dataset. Finally, there were 2450-K unique selected tweets.

### 3.3 Labeling

In this study, four annotators with knowledge of the meanings of various labels and classes assisted in the labelling procedure. The determined labels include eleven distinct categories that denote the nature of each tweet and its target recipient, including Non-Offensive (0-NO), Offensive (1-Offensive), and Hate (2-Race & Ethnicity, 2-Nationality, 2-Age, 2-Sex/Gender, 2-Guild, 2-Disability, 2-Immigration Status, 2-Religion, and 2-Victims of major violent event). Although we do not need to know the target group of each hate tweet to accomplish the research objective of detecting Hate and Offensive speech, we annotated the tweets with target group specifications to assess and improve the accuracy, prevent data bias, and use the annotated dataset in future research (such as detecting target group of hate

---

<sup>2</sup> TWINT.

**Table 1** Data statistics per label

Label	Count
0- NO	18,626
1- Offensive	16,349
2- Religion	1970
2- Nationality	376
2- Race and ethnicity	320
2-Guild	223
2- Sex/gender	98
2- Age	20
2- Disability	10
2- Immigration status	7
2- Victims	1
Total	38,000

speech). However, for the sake of the statistical research, we consider all labels beginning with the "2" prefix to be Hate.

The total number of tweets selected for labelling is 38-K, and it consists of various combinations of tweets with varied settings (the presence or absence of offensive keywords and keywords associated with target groups) that were randomly selected in each configuration. The labelling process began with a train to ensure that annotators shared a common understanding and perception of the various label definitions. Thus, a small dataset of 1000 randomly selected tweets was annotated independently by each annotator. The outcomes of each annotator's labels were compared to those of the others using Cohen's Kappa Coefficient (a measure of inter-annotator reliability with a range of  $-1$  to  $1$ ) (Cohen 1960). The average outcome of the comparison is 0.7596, indicating that the annotators pay careful attention to the definition and labels.

In the following, The tweets were randomly divided into three subsets, which two random and independent annotators labelled. Finally, each tweet contains two labels; if both labels were in the same class (0- NO, 1- Offensive, 2- Hate), one of the labels was given to the tweet's class; if not, the tweet's class was selected by group consensus.

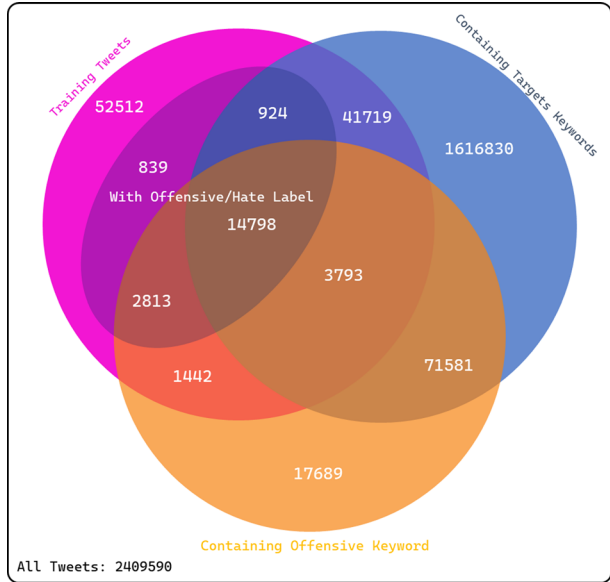
### 3.4 Data statistics

Table 1 summarizes the number of annotated tweets in each class. The tweet counts in some categories are low which indicates in Persian, based on the dominant culture of Persian speakers, insulting these groups is not common. Figure 1 depicts the final statistics results. The total number of tweets, tweets containing target keywords, and tweets containing offensive terms is around 2409-K, 1749-K, and 112-K, respectively. More detailed statistics can be seen in Fig. 1.

## 4 Reducing sampling bias

Machine learning models are trained to somehow make decision based on bias in data (Dixon et al., 2018). Offensive content detection models are also biased in favor of offensive samples, giving them a higher score than neutral samples. Model prediction, on the



**Fig. 1** Data distribution statistics

other hand, should not be influenced by factors such as gender, race, religion, age, etc. Otherwise, we risk encountering unintended bias. Due to the breadth of different aspects of unintended bias, examining it in practical examples is generally difficult. The less unintended bias occurs in the models, the more efficient they are (Dixon et al., 2018). When we refer to reducing bias, we mean reducing unintended bias. The methods proposed so far are also intended to somehow reduce unintended bias.

Sampling bias in the dataset building process will lead to unintended bias in model prediction. Offensive tweets account for only 3% of all the tweets on Twitter (Fortuna and Nunes, 2018). According to our study, offensive tweets make up approximately 3.5% of all Persian tweets. Therefore, by random sampling, a very significant number of tweets must be tagged to obtain a sufficient amount of offensive data. To tackle this problem, one approach is to use keyword-based sampling to achieve a higher number of offensive tweets. However, using a keyword-based sampling method may result in a strong correlation between specific keywords and the offensive label (Dixon et al., 2018; Garg et al., 2022). In the following sections, we discuss the bias evaluation criteria and how to calculate it on the collected dataset.

#### 4.1 Bias evaluation metrics

Bias evaluation metrics False Positive Equality Difference (FPED) and Pinned AUC Equality Difference (pAUCED) are defined in (Dixon et al., 2018). FPED is a measure of variation in opportunity equality and calculated based on variations of term-wise error rates ( $FPR_t$ ) around the error rates ( $FPR$ ) of the complete evaluation set:

$$FPED_T = \sum_{t \in T} |FPR - FPR_t|. \quad (1)$$

Every term-wise error rate can be between zero and one, so the sum of the above-mentioned relation can be between 0 and number of terms we want to study. Any value close to zero indicates less bias in the dataset, and vice versa.

FPED is threshold dependent and require a classifier that produces binary labels. Pinned AUC ( $pAUC_t$ ) developed for a term  $t$ , which is the AUC measure on a pinned dataset  $pD_t$ , such that  $pD_t = s(D_t) \cup s(D)$  and  $|s(D_t)| = |s(D)|$ , where  $s(D_t)$  is the set of documents containing the term  $t$  in the evaluation set,  $s(D)$  is the complete evaluation set, and  $s(\cdot)$  is a sampling function:

$$pAUCED_t = \sum_{i \in T} |AUC - pAUC_t|, \quad (2)$$

where AUC is calculated on the complete test set.

False positive equality differences can only be used to assess bias in the context of direct binary classification or after a threshold has been set. The pinned AUC metric offers a threshold-agnostic approach for detecting bias in a broader range of usecases (Dixon et al., 2018).

## 4.2 Bias analysis in dataset

Initially, experts introduced issues on which we believe the data should not be biased. They also proposed a handful of words selected from the offensive and hate speech list of keywords as potential candidates for bias study. There are 268 words in this set, covering issues including religion, politics, gender discrimination, ethnicity, physical traits, and miscellaneous. In parallel, a Transformer-based model BertTweet-FA (see section 5) is fine-tuned on the original dataset, and the test set labels are predicted. Following that, based on the prediction results of the test data, the FPED measure is calculated for each of the candidate words, and the words with this measure greater than 0.15 are picked as identity keywords. This results in a collection of 51 identity keywords, and the tweets containing these words are misclassified at a high rate. To see the translation of these words (using Google Translate) and transliteration, refer to Appendix A.

To determine whether we might be biased toward these keywords in the training dataset, we calculated the frequency of the identity keywords in offensive tweets and across all the datasets. Figure 2 shows that in almost all cases, the frequency of these words in offensive tweets is higher than the total number of tweets. So, given the same settings, it appears that models can be expected to have a more negative perception of tweets containing these terms, as contrasted to identical tweets without these words. It should be noted that none of these identity keywords are considered insults, but many controversial tweets contain them. Models that are trained on such a dataset would become more sensitive to such words, resulting in an increase in false positives.

## 4.3 Debiasing the dataset

To alleviate unintended bias due to keyword-based sampling, we have added additional not-offensive data to address the imbalance between not-offensive and offensive tweets corresponding to 51 identity keywords.

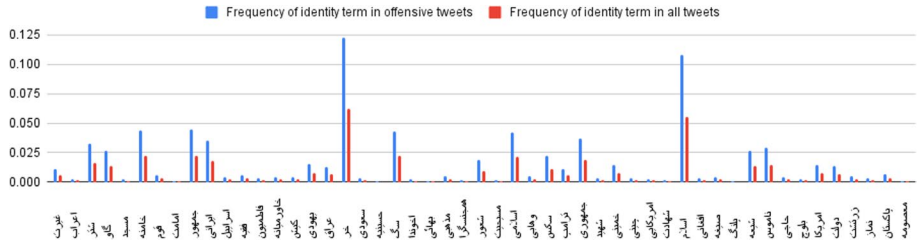


Fig. 2 Frequency of identity keywords in offensive tweets and all dataset

The method of adding data is that for each keyword, tweets containing these words but not offensive or hateful keywords are filtered from the initial set of Persian tweets (excluding the original dataset tweets). The collected tweets are considered not-offensive due to the low number of offensive tweets as well as filtering with a comprehensive set of the offensive and hateful lexicon. In this way, without the use of human labeling, the data is balanced against identity keywords. We are careful not to upset the balance of other identity keywords when we add not-offensive tweets containing one specific word during this process.

### 5 Experimental setup

Two factors significantly impact the results in the field of Persian offensive language detection: selecting and altering the type and volume of data in the training and test datasets and selecting, implementing, and configuring the parameters of NLP models to identify classes. The following information and configurations pertain to the test and training data and the applied models and their configurations:

*Training and test datasets:* To produce the dataset, we first removed data from classes containing fewer than 100 tweets, such as Sex, Age, Disability, Immigration Status, and Victims. The final count of the judged labeled data is 37,868. The number of tweets in the non-offensive class (18,626) is nearly equivalent to the combined total number of tweets in the offensive and hate classes (19,238). Stratified sampling was used to select 5000 tweets from the dataset as test data. This included 2540 tweets from the Hate and Offensive classes, with the same proportional representation as the whole dataset, as well as 2460 non-offensive tweets.

*Models:* We explored different machine learning methods commonly used in previous works on offensive language detection such as Support Vector Machine (SVM) (Davidson et al., 2017; Salminen et al., 2018), Logistic Regression (LR) (Zampieri et al., 2019) as well as Transformer-based models, which have recently gained popularity in NLP community. We finetuned mBERT (Devlin et al., 2019), ParsBert (Farahani et al., 2021), XLM-RoBERTa (Conneau et al., 2019) and BERTweet-FA (Malekzadeh, 2020) (BERT model pre-trained on 20 milion Persian tweets) on our dataset. The reasons for choosing BERT-based language models are that they are effective for classification because they are pre-trained on a large quantity of text data, capture contextual relationships between words, are bidirectional, and can be fine-tuned for specific tasks with minimal additional data. They have attained state-of-the-art performance on numerous NLP benchmark tasks, making them a popular classification option (Wu et al. 2022; Arslan et al. 2021; Barbieri et al.

2020). All these models are used in base and uncased form. In the following, the hyperparameters of the models are reported:

1. SVM: C-Support Vector Classification with "linear" kernel and 1.0 regularization. The degree of the polynomial kernel is 3.
2. LR: With "5e1" inverse of regularization and "lbfgs" algorithm to optimize the problem. The loss minimized is the multinomial loss fit across the entire probability distribution. Also, the random state is set to 17.
3. Transformer-based: batch size=64, learning rate=3e-6, epoch=6.

## 6 Results and discussions

In the following, the results of the experiments performed in this research are presented and analyzed.

### 6.1 Persian offensive language detection

We conduct an experiment to investigate the accuracy of different models in detecting offensive content in Persian. We apply our analysis on two types of models: (1) mostly used traditional ML classifiers include SVM and LR using bag of words features, and (2) fine-tuned PLMs include mBERT, XLM-R, ParsBERT and BERTweet-FA. PLMs insert a specific CLS token at the start of the input sentence and feed it into stacked layers of Transformer encoders. The final layer's representation of the CLS token is sent into a linear layer for 2-way classification (Offensive or not-offensive).

Performance of the models are reported in Table 2. As expected, according to experiments on the original dataset, pre-trained language models gained higher performance. This is because these models are very strong in learning contextual relations between words. Most misclassified samples in SVM and LR predictions are tweets that do not necessarily contain an offensive keyword but are truly offensive, whereas these errors are lower in Transformer-based models, indicating that the models tend to classify an instance as belonging to the offensive class even if it does not explicitly contain offensive terms.

**Table 2** Precision, recall, *F1* on Persian test set

	Original dataset			Debiased dataset		
	P	R	F1	P	R	F1
SVM	0.857	0.857	0.857	0.85	0.86	0.855
LR	0.851	0.851	0.851	0.857	0.852	0.857
mBERT	0.893	0.892	0.892	0.894	0.895	0.895
ParsBERT	0.864	0.863	0.863	0.862	0.861	0.861
XLM-RoBERTa	0.894	0.893	0.893	0.896	0.897	0.897
BERTweet-FA	<b>0.903</b>	<b>0.903</b>	<b>0.903</b>	<b>0.902</b>	<b>0.902</b>	<b>0.902</b>

We report the values for models trained with original and debiased dataset

Bold shows that in each column that number which is the highest score

The BERTweet-FA model performed best and achieved 0.915 in F1-score in this experiment. The reason is that the model trained on a large Persian tweet dataset and its domain is quite similar to the domain of our dataset.

Despite the fact that the ParsBERT is a BERT model which is trained on a large Persian corpus, it performed poorly in comparison to other Transformers. This model appears to have been trained with less data than XLM-RoBERTa and m-BERT. Also, the data domain on which this model is trained is different from the BERTweet-FA model, which is trained on Persian tweet dataset. BERTweet-FA may also have a more precise tokenizer for the proposed dataset. As a result, it seems that ParsBERT is less accurate to encode a tweet than other Transformers employed in this study.

According to (Fortuna et al., 2021; Zampieri et al., 2020), when compared to the results of state-of-the-art models on datasets of other languages, our results on Persian are comparable and even superior.

## 6.2 Debiasing analysis

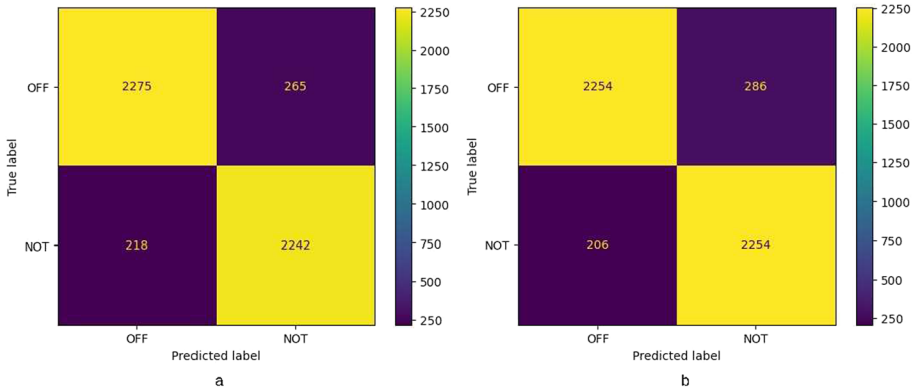
To demonstrate the effects debiasing data, we present the performance of classifiers trained on the Persian dataset. According to Table 2, the model results on bias reduced data are largely comparable to those trained on original data. This shows that after reducing bias toward identity keywords in the data, the models almost retain their performance.

Table 3 show the bias measurement experiment on our dataset. We only consider Transformer-based models for this experiment due of their higher F1-score. The sum pAUCED for Transformer-based models on the test set and based on identity keywords, before and after debiasing data is shown in the Table 3. As expected, in most models the sum FPED and pAUCED decreased after the bias was reduced in terms of identifier words. These results on the two measures and results on Table 2 demonstrates a reduction in unintended bias without sacrificing general model performance. According to FPED result in Table 3, the maximum drop on FPED measure happened for BERTweet-FA model. FPED and pAUCED have increased for ParsBERT and BERTweet-FA, respectively. Despite the fact that we contributed data to address the bias for identity keywords, it seems that this has caused a different type of bias. Thus, these bias measurement criteria have increased since the estimation of these models on the test data has changed, and as a result, one of the parameters for computing these criteria, the average on the whole test data, has changed.

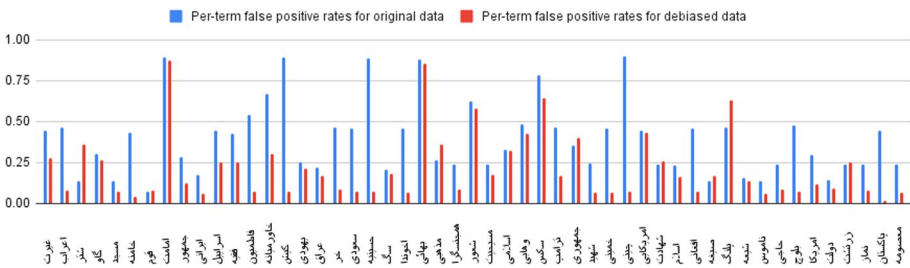
Figure 3a and b display the confusion matrices of BERTweet-FA model against the test set data. It can be seen that after bias reduction in train data, false positives increased and false negatives decreased. Despite the increase in false positives, the model's performance has not dropped in terms of F1-score. What matters is that the number of false negatives has decreased, indicating that the model recognizes more offensive tweets. Because the

**Table 3** Sum pAUCED and FPED over identity keywords in original and debiased settings

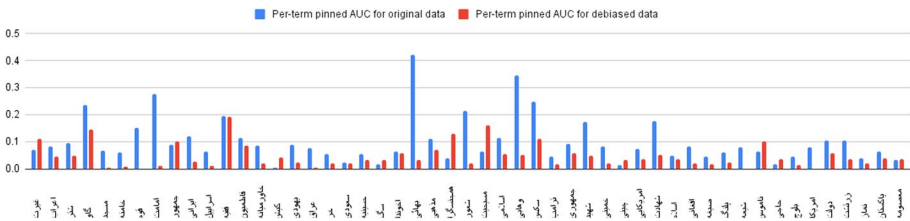
	FPED		pAUCED	
	Original	Debiased	Original	Debiased
mBERT	13.82	10.76	7.84	5.12
ParsBERT	11.41	9.21	5.42	5.78
XLM-RoBERTa	11.45	13.56	3.44	2.34
BERTweet-FA	14.30	11.21	3.53	2.11



**Fig. 3** Confusion matrix of the best experiments (BERTweet-FA) for the Persian language **a** original train data, and **b** bias reduced train data



**Fig. 4** Per-term false positive rates for original and debiased dataset



**Fig. 5** Per-term pinned AUC for original and debiased dataset

spread of offensive content promotes violence against certain minorities (Fortuna and Nunes, 2018), it is desirable to detect as much of this content as possible, although it may lead to the blocking of more neutral tweets (Wullach et al., 2021).

The FPED and pAUCED values for each word were calculated using the BERTweet-FA model before and after debiasing on the test data, and are shown in Figs. 4 and 5, respectively. These two measures have decreased over most words, indicating a decrease in bias towards the identity keywords.

**Table 4** Precision, recall, F1 on Persian test set in imbalanced data settings

	Original data		
	P	R	F1
Mild	0.903	0.903	0.903
Moderate	0.891	0.892	0.892
Extreme	0.83	0.713	0.767

### 6.3 Imbalanced data setting

We establish an experiment to assess the model's performance on this dataset in the presence of data imbalance. We investigate three conditions of imbalanced data settings: (1) mild, (2) moderate, and (3) extreme in which the proportion of the minority class, in our study the offensive label, is 20-40%, 1-20%, and <1% of the dataset, respectively. We use BERTweet-FA to run the experiment. According to the results of Table 4, the model performs well in terms of mild and moderate imbalanced data settings.

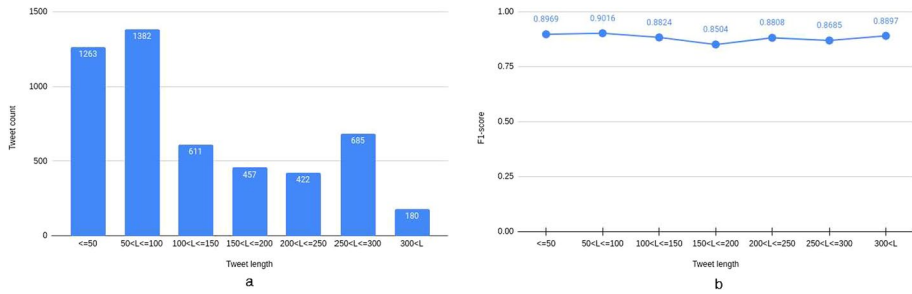
Given that the dataset is collected over one year, covers various events, and includes offensive tweets in different forms and expressions, it seems that even in unbalanced data settings, the variety of offensive tweets is wide enough, making the model more resistant to performance reduction. In addition, the power of Transformer models in generalizability is another reason for the robustness of the model against the imbalanced data configuration. In extreme imbalance condition, 0.75 in the F1-score is obtained only by considering 200 offensive tweets in train data, which is acceptable.

### 6.4 Robustness towards the variability of tweet length

In this section, we conduct an experiment to investigate the sensitivity to tweet length. According to our study, an insult in Persian can be expressed in a limited number of words, so it is important to examine the model's performance by changing the length of the tweets. We use the number 50 as the offset in this experiment. We consider tweets with a length of less than a certain number of characters based on this offset from the test set each time and calculate the model's F1-score on this set. According to the results shown in Fig. 6, the model's F1-score is well maintained on tweets of varying lengths, indicating that the model also has acceptable performance for data of shorter lengths.

## 7 Conclusion and future work

We gathered a Persian tweet dataset for offensive language detection. Because labeling is an expensive process, keyword-based sampling methods were used to overcome the problem of data imbalance. We investigated the impact of data bias on offensive language detection and discovered that this issue is closely related to how data is sampled. Then we examined the candidate words that may introduce bias into the dataset, and by adding data, we attempted to remove the bias toward this set of words, which ultimately reduced the overall bias of the data.



**Fig. 6** Length sensitivity analysis: **a** tweets count, and **b** F1-score according to tweet length

We applied various classifications of common machine learning methods and Transformer-based models to the proposed dataset, and experiments show that Transformer-based models detect offensive content more efficiently. Increasing the number of instances of the proposed dataset will be considered in the future. The greater the diversity of offensive tweets in the dataset, the better it will be for training models. Improving preprocessing modules, such as tokenizers and emoji-to-text models in the Persian language, can also improve performance. Furthermore, a closer look at debiasing the models over the data can be effective in improving performance. The human step of identifying the bias-prone identity terms can be eliminated in future works. Although our work is preliminary on the proposed dataset, we hope that a path has been taken to examine the Persian offensive language and its various aspects.

## Appendices

### Identity Keywords

Full list of identity keywords in Persian, English translation(using Google Translate), and with transliteration are shown in Fig. 7.



Persian	English Translation	Transliteration	Persian	English Translation	Transliteration
غیرت	Jealousy	Qeyrat	مذهبی	Religious	Mazhabi
اعراب	Arabs	A'raab	همجنسگرا	Homosexual	Hamjensgera
شتر	Camel	Shotor	شعور	Consciousness	Shour
گاو	Cow	Gav	مسیحیت	Christianity	Masihiat
مسجد	Mosque	Masjed	اسلامی	Islamic	Eslami
خامنه	Khamenei	Khamenei	وهابی	Wahhabi	Wahhabi
قوم	Race	Qom	سکس	Sex	Sex
امامت	Imamate	Emamat	ترامپ	Trump	Trump
جمهور	Republic	Jomhoor	جمهوری	Republic	Jomhoori
ایرانی	Iranian	Irani	شهید	Martyr	Shahid
اسرائیل	Israel	Esraeil	خمینی	Khomeini	Khomeini
فقیه	Faqih	Faqih	چینی	Chinese	Chini
فاطمیون	Fatimid	Fatemiyoun	امریکایی	American	Amrikaei
خاورمیانه	Middle East	Khavarmiane	شهادت	Witness	Shahadat
کیش	Kish	Kish	اسلام	Islam	Eslam
یهودی	Jewish	Yahoudi	افغانی	Afghan	Afghani
عراق	Iraq	Araq	صیغه	Concubine	Sighe
خر	Donkey	Khar	پلنگ	Leopard	Palang
سعودی	Saudi	Soudi	شیعه	Shia	Shie
حسینیه	Hosseinieh	Hosseinieh	ناموس	Honor	Namoos
سگ	Dog	Sag	حاجی	Pilgrim	Haji
اخوندا	Mullah	Akhunda	بلوچ	Baloch	Balooch
بهائی	Baha'i	Baha'i	امریکا	America	Amrika
دولت	Government	Dolat	نماز	Prayer	Namaz
زرتشت	Zoroaster	Zartosht	پاکستان	Pakistan	Pakestan
معصومه	Massoumeh	Massoumeh			

**Fig. 7** Identity keywords in Persian, English (using Google Translate), and transliteration

**Acknowledgements** This research was in part supported by a grant from the School of Computer Science, Institute for Research in Fundamental Sciences, IPM, Iran (No. CS1402-4-237).

**Author Contributions** All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Emad Kebriaei, Ali Homayouni, Roghayeh Faraji and Armita Razavi. The first draft of the manuscript was written by Emad Kebriaei. Reviewing the manuscript and editing is done by Azadeh Shakery, Hesham Faili and Yadollah Yaghoobzadeh. All authors read and approved the final manuscript.

**Funding** This research was in part supported by a grant from the School of Computer Science, Institute for Research in Fundamental Sciences, IPM (No. CS1400-4-237).

**Availability of data and materials** The data that support the findings of this study are available from the corresponding author, upon reasonable request.

## Declarations

**Conflict of interest** We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

**Ethics approval** We confirm that the manuscript would not be submitted for publication in any other Journal or Magazine till the decision is made by journal editors.

**Consent to participate** We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

**Consent for publication** Not applicable.

**Code availability** Code for data cleaning and analysis is provided as part of the replication package. It is available at <https://www.dropbox.com/s/z09fjb84wcaqqvn/HSD.zip?dl=0> for review.

## References

- (2019). Jigsaw unintended bias in toxicity classification. <https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/>
- Alavi, P., Nikvand, P., & Shamsfard, M. (2021). Offensive language detection with bert-based models, by customizing attention probabilities. CoRR [arXiv:abs/2110.05133](https://arxiv.org/abs/2110.05133).
- Aldjanabi, W., Dahou, A., Al-qaness, M. A., et al. (2021). Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. In *Informatics, Multidisciplinary Digital Publishing Institute*, p. 69.
- Aljarah, I., Habib, M., Hijazi, N., et al. (2021). Intelligent detection of hate speech in arabic social network: A machine learning approach. *Journal of Information Science*, 47(4), 483–501.
- Aljero, M. K. A., & Dimililer, N. (2021). A novel stacked ensemble for hate speech recognition. *Applied Sciences*, 11(24), 11,684.
- Alshalan, R., Al-Khalifa, H., Alsaeed, D., et al. (2020). Detection of hate speech in covid-19-related tweets in the Arab region: Deep learning and topic modeling approach. *Journal of Medical Internet Research*, 22(12), e22,609.
- Arslan, Y., Allix, K., Veiber, L., et al. (2021). A comparison of pre-trained language models for multi-class text classification in the financial domain. *Companion Proceedings of the Web Conference, 2021*, 260–268.
- Badjatiya, P., Gupta, S., Gupta, M., et al. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pp. 759–760.
- Badjatiya, P., Gupta, M., & Varma, V. (2019). Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, pp. 49–59.
- Barbieri, F., Camacho-Collados, J., Neves, L., et al. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. arXiv preprint [arXiv:2010.12421](https://arxiv.org/abs/2010.12421)
- Basile, V., Bosco, C., Fersini, E., et al. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th international workshop on semantic evaluation, association for computational linguistics*, pp. 54–63.
- Chiu, K. L., & Alexander, R. (2021). Detecting hate speech with gpt-3. [arXiv:2103.12407](https://arxiv.org/abs/2103.12407)
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Conneau, A., Khandelwal, K., Goyal, N., et al. (2019). Unsupervised cross-lingual representation learning at scale. [arXiv:1911.02116](https://arxiv.org/abs/1911.02116)
- Czarnowska, P., Vyas, Y., & Shah, K. (2021). Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9, 1249–1267.
- Davidson, T., Warmusley, D., Macy, M., et al. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, pp. 512–515.
- Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. [arXiv:1905.12516](https://arxiv.org/abs/1905.12516)
- Dehghani, M., Dehkordy, D. T., & Bahrani, M. (2021). Abusive words detection in persian tweets using machine learning and deep learning techniques. In *2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, IEEE (pp. 1–5).
- Devlin, J., Chang, M. W., Lee, K., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, Volume 1 (Long


- and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota (pp. 4171–4186). <https://doi.org/10.18653/v1/N19-1423>, <https://aclanthology.org/N19-1423>
- Dixon, L., Li, J., Sorensen, J., et al. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society* (pp. 67–73).
- Dowlagar, S., & Mamidi, R. (2021). Hasocone@ fire-hasoc2020: Using bert and multilingual bert models for hate speech detection. [arXiv:2101.09007](https://arxiv.org/abs/2101.09007)
- Gharachorloo, M., Farahani, M., Farahani, M., et al. (2021). Parsbert: Transformer-based model for Persian language understanding. *Neural Processing Letters*, 53(6), 3831–3847.
- Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4), 1–30.
- Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing and Management*, 58(3), 102,524.
- Founta, A. M., Djouvas, C., Chatzakou, D., et al. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth international AAAI conference on web and social media*
- Garg, T., Masud, S., Suresh, T., et al. (2022). Handling bias in toxic speech detection: A survey. [arXiv:2202.00126](https://arxiv.org/abs/2202.00126)
- Golbeck, J., Ashktorab, Z., Banjo, R. O., et al. (2017). A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*. Association for Computing Machinery, New York, NY, USA, WebSci '17, (p. 229–233), <https://doi.org/10.1145/3091478.3091509>
- Haq, N. U., Ullah, M., Khan, R., et al. (2020). Usad: An intelligent system for slang and abusive text detection in Perso-Arabic-Scripted Urdu. *Complexity* 2020.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29.
- He, B., Ziems, C., Soni, S., et al. (2021). Racism is a virus: anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, (pp. 90–94).
- Jey, P. S., Hemmati, A., Toosi, R., et al. (2022). Hate sentiment recognition system for persian language. In *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, IEEE, (pp. 517–522).
- Kennedy, B., Jin, X., Davani, A. M., et al. (2020). Contextualizing hate speech classifiers with post-hoc explanation. [arXiv:2005.02439](https://arxiv.org/abs/2005.02439)
- Kennedy, G., McCollough, A., Dixon, E., et al. (2017). Technology solutions to combat online harassment. In *Proceedings of the first workshop on abusive language online*. Association for Computational Linguistics, Vancouver, BC, Canada, (pp. 73–77), <https://doi.org/10.18653/v1/W17-3011>, <https://aclanthology.org/W17-3011>
- Madukwe, K., Gao, X., & Xue, B. (2020). In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the fourth workshop on online abuse and harms*, pp. 150–161.
- Malekzadeh, A. (2020). Bertweet-fa: A pre-trained language model for persian (a.k.a farsi) tweets. <https://github.com/arm-on/BERTweet-FA>
- Mollas, I., Chrysopoulou, Z., Karlos, S., et al. (2020). Ethos: An online hate speech detection dataset. [arXiv:2006.08328](https://arxiv.org/abs/2006.08328)
- Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A bert-based transfer learning approach for hate speech detection in online social media. In *International conference on complex networks and their applications*, Springer, (pp. 928–940).
- Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on bert model. *PLoS One*, 15(8), e0237,861.
- Mozafari, M., Farahbakhsh, R., & Crespi, N. (2022). Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10, 14,880–14,896. <https://doi.org/10.1109/ACCESS.2022.3147588>
- Park, J. H., Shin, J., & Fung, P. (2018). Reducing gender bias in abusive language detection. [arXiv:1808.07231](https://arxiv.org/abs/1808.07231)
- Qian, J., Bethke, A., Liu, Y., et al. (2019). A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, (pp. 4755–4764), <https://doi.org/10.18653/v1/D19-1482>, <https://aclanthology.org/D19-1482>
- Rajput, G., Punn, N. S., Sonbhadra, S. K., et al. (2021). Hate speech detection using static bert embeddings. In *International conference on big data analytics*, Springer, (pp. 67–77).

- Salawu, S., He, Y., & Lumsden, J. (2020). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, 11(1), 3–24. <https://doi.org/10.1109/TAFFC.2017.2761757>
- Salminen, J., Almerikhi, H., Milenković, M., et al. (2018). Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Twelfth International AAAI Conference on Web and Social Media*.
- Sap, M., Gabriel, S., Qin, L., et al. (2019). Social bias frames: Reasoning about social and power implications of language. [arXiv:1911.03891](https://arxiv.org/abs/1911.03891)
- Schmidt, A., & Wiegand, M. (2019). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, April 3, 2017, Valencia, Spain, Association for Computational Linguistics, (pp. 1–10).
- Shah, D., Schwartz, H. A., & Hovy, D. (2019). Predictive biases in natural language processing models: A conceptual framework and overview. [arXiv:1912.11078](https://arxiv.org/abs/1912.11078)
- Silva, L., Mondal, M., Correa, D., et al. (2016). Analyzing the targets of hate in online social media. In *Tenth international AAAI conference on web and social media*.
- Van Hee, C., Lefever, E., Verhoeven, B., et al. (2015). Detection and fine-grained classification of cyberbullying events. In *Proceedings of the international conference recent advances in natural language processing*, (pp. 672–680).
- Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, (pp. 88–93).
- Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of abusive language: The problem of biased datasets. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, (pp. 602–608). <https://doi.org/10.18653/v1/N19-1060>, <https://aclanthology.org/N19-1060>
- Wu, T., Caccia, M., Li, Z., et al. (2022). Pretrained language model in continual learning: A comparative study. In *International conference on learning representations*.
- Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web. International world wide web conferences steering committee, republic and canton of Geneva, CHE, WWW '17*, (p. 1391-1399), <https://doi.org/10.1145/3038912.3052591>
- Wullach, T., Adler, A., & Minkov, E. (2021). Towards hate speech detection at large via deep generative modeling. *IEEE Internet Computing*, 25(2), 48–57. <https://doi.org/10.1109/MIC.2020.3033161>
- Zampieri, M., Malmasi, S., Nakov, P., et al. (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). [arXiv:1903.08983](https://arxiv.org/abs/1903.08983)
- Zampieri, M., Nakov, P., Rosenthal, S., et al. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the fourteenth workshop on semantic evaluation. International committee for computational linguistics, Barcelona (online)*, (pp. 1425–1447), <https://doi.org/10.18653/v1/2020.semeval-1.188>, URL <https://aclanthology.org/2020.semeval-1.188>
- Zhang, C., Beetz, J., & de Vries, B. (2018). Bimsparql: Domain-specific functional sparql extensions for querying rdf building data. *Semantic Web*, 9(6), 829–855.
- Zhou, X. (2021). *Challenges in automated debiasing for toxic language detection*. University of Washington.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

Emad Kebriaei<sup>1</sup>  · Ali Homayouni<sup>1</sup> · Roghayeh Faraji<sup>1</sup> · Armita Razavi<sup>1</sup> · Azadeh Shakery<sup>1,2</sup> · Heshaam Faili<sup>1,2</sup> · Yadollah Yaghoobzadeh<sup>1</sup>

✉ Emad Kebriaei  
emad.kebriaei@ut.ac.ir

Ali Homayouni  
alihomayouni@ut.ac.ir

Roghayeh Faraji  
roghaye.farajiii@gmail.com

Armita Razavi  
armita.razavi@gmail.com

Azadeh Shakery  
shakery@ut.ac.ir

Heshaam Faili  
hfaili@ut.ac.ir

Yadollah Yaghoobzadeh  
y.yaghoobzadeh@ut.ac.ir

<sup>1</sup> School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

<sup>2</sup> School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran