Tabriz Journal of Electrical Engineering (TJEE), vol. 53, no. 2, Summer 2023 Serial no. 104
DOI: 10.22034/tjee.2023.15794

PerBOLD: A Big Dataset of Persian Offensive language on Instagram Comments

Maryam Khodabakhsh[1], Fateme Jafarinejad[1*], Marziea Rahimi[1], Masood Ghayoomi[2]

[1]Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran.
[2]Institute for Humanities and Cultural Studies, Ghom, Iran.
m_khodabakhsh@shahroodut.ac.ir, jafarinejad@shahroodut.ac.ir, marziea.rahimi@shahroodut.ac.ir,
M_Ghayoomi@ihcs.ac.ir
[*]Corresponding author

Received 19/09/2022
Accepted 26/12/2022.

Abstract
Easy access to social media enables users to express their opinions and ideology about various topics like news, videos, and personalities freely, without any fear, and often in an offensive manner. It is a vital task to detect comments with offensive language on social media platforms and relies on a complete and comprehensive tagged dataset. Therefore, in this paper, we introduce and make publicly available PerBOLD, a new Persian comment dataset collected from Instagram as a popular platform among Iranian. We follow a two-level manual annotation process in order to determine whether a comment has offensive language or not and fine-grained tags of different types of offensive language. Furthermore, we present some interesting aspects of data and analysis them.

Keywords
Natural language processing, offensive language, social media, annotation.

1. Introduction

With the expansion of Internet usage, social networks provided an unprecedented opportunity for interaction of users to share information, daily activities, and conflicting opinions. Of course, the development of technology, like a double-edged sword, may also be associated with some risks. Increasing the use of Socially Unaccepted Discourse (SUD) and violations of related laws are among the most important risks and challenges. The relative freedom of cyberspace, the relative anonymity of users' identities, the low level of media literacy, and the lack of effective regulations provided by social networking platforms can be the most important causes of this phenomenon. However, the speed of information production in social networks and the web is so high that the use of manual methods in detecting and filtering SUD is practically inefficient. Therefore, many researchers utilize from natural language processing (NLP) algorithms for automatic identification. NLP expands its applications in wide range of fields e.g. classifying movie genre, analysing social networks, automatic keyword generation [1, 2]..

Detection of Hostility [3-6], Cyber bullying [7, 8], hate speech [3, 9-14], and offensive speech [15-17] are among the efforts made in the field of automatic detection of some categories of SUD. In these researches, various types of texts including news texts [4] and user posts or comments on various social networks such as Facebook [18], Twitter [5-7, 11, 15], Reddit [16], Gap [9], and Instagram [8, 19] have been considered. These studies have been done in different languages such as English [3, 4, 8, 16], Arabic [3], German [17], Spanish [11], Italian [7], and a combination of several Language [13].

Offensive speech is one of the most important types of SUD, various researches have been conducted on the introduction of Offensive Language Identification Dataset (OLID) and automatic identification of offensive language on texts. According to the [3], offensive speech can be defined as speech that can be harmful to readers (e.g. containing humiliating, and insulting words). According to [20], offensive Language is commonly defined as hurtful, derogatory, or obscene comments made by one person to another person. Qian et.al. [9] has considered language containing toxic, hateful, abusive, violent, and bullying as the characteristics of offensive language. It tagged offensive (non-offensive) posts of the OffenseEval dataset [15] as posts including insults, threats, and posts containing any form of untargeted profanity (excluding all of these).

Due to the limitations of Persian language resources, few researches have been done on the automatic detection of any type of SUD in this language. Among these researches [21], [22] focused on identification of offensive languages on Persian data as a low-resource language. Alavi et al. [21] Construct a dataset of about 5k Persian text data comprising 2,453 inoffensive and 2,535 offensive data. Mozafari [22] provide another one-level tagged OLID dataset of offensive data consisting of 6k micro blog posts from Twitter.

As another work of identification of offensive language on low-resource languages, this paper will present a large collection of tagged data (about 30k data) of offensive texts in the Persian language, which is much

larger than the existing Persian OLID datasets. All these data are manually tagged. Another innovation of this dataset is in providing very fine-grained tags of different types of offensive speech in a two-level tagging approach. The other existing Persian OLID datasets provide an one-level tagging ([21]) or provide just a coarse-grained tagging of targeted/untargeted or individual/group classification of offensive comments ([22]). To prepare this dataset, a combination of user-based and news agency-based approaches has been used. The comments of Instagram users to the posts of some specific users or news agencies have been crawled. These users were selected from social or political celebrities. In the case of news agencies, some well-known news agencies that had news with different social, cultural, economic, and political genres were selected. To the best of our knowledge, the Instagram news agency-based approach has not been used so far in SUD datasets.

The structure of this paper is as follows: In Section 2, we will review the research done around the production of offensive speech datasets or the automatic detection of offensive speech. In the third section, we will describe the procedure performed in creating the PerBOLD dataset. The fourth section describes the statistical characteristics of the PerBOLD dataset. Finally, the fifth section will present the conclusion.

2. Related works

In this section, we first review prior work on offensive language datasets from social media comments and then present the related literature on offensive datasets in the Persian language.

2.1. Offensive Language Datasets

There is a rising use of offensive language on social media platforms like YouTube [23-25], Facebook [24, 25], and Twitter [26-28] in recent years. It is important to detect and remove such use of offensive content automatically for many websites and organizations. The task of automatic detection of offensive language from social media has attracted the attention of NLP community researchers its success depends on the complete and comprehensive tagged dataset. Therefore, our interest in this paper is the collections annotated for offensive language and we provide a review focusing on offensive language datasets, the approaches of collecting text data, and their annotation processes.

A famous approach to build an offensive dataset is using abusive words as references to collect comments. For example, researchers in [27], retrieved the tweets based on searching for words and constructions that are often included in offensive messages using Twitter API and introduced the OLID. Tweets in the dataset are annotated for offensive content using a three-level hierarchy annotation scheme where a tweet is labeled as offensive (OFF) if it contains any form of profanity or targeted offense, and non-offensive (NON) otherwise. The training part of the OLID dataset contains 13,241 samples, while the testing part contains 860 tweets. Díaz-Torres et al. [28] built a corpus of 10,500 tweets from August to November 2017 in Mexican Spanish using some rude words and controversial hashtags. The main focus of their research was on the annotation process which provided a specific criterion to separate a tweet from aggressive, offensive, and vulgar, based on the linguistic characteristics and intent of the message. Be any of these classes, the tweet would be labelled as offensive.

Addition to retrieve tweets based on searching the offensive words and controversial hashtags, researchers in [26] benefited from other methods to collect tweets. For example, they looked for posts that are defending a particular group as replay tweets, and then followed those tweets to the original ones. The latter is usually an instance of hate speech. Also, based on the fact that certain Twitter users such as celebrities are more likely to be targeted by both hateful and offensive speech, they identified the list of people who could be a potential target of offensive and hate speech and then used the Twitter API to retrieve tweets that mention those user accounts. After doing cleaning operations, their corpus included 5361 tweets. Each tweet was manually annotated by three experts whether it contains offensive language or not.

Searching based on a specific topic is another approach to build a dataset. For instance, Kannada CodeMixed Dataset (KanCMD) [23] consists of Youtube comments from 18 videos on different topics ranging from movie trailers to current trends about the ban on mobile apps in India, India-China border issue, Mahabharata, and Transgenders. The label of each comment was determined manually by a minimum of three annotators and a maximum of five annotators. In another recent study [29], Hada et al. presented a dataset of 6000 English language Reddit comments. Based on the fact that Reddit contains forums called subreddits dedicated to specific topics and allows users to make a post on the subreddit to start a discussion, they chose subreddits to cover a diverse range of topics from generic themes to controversial. Also, they collected comments from random subreddits. They identified the label of each comment by four annotators using Crowdsourcing of Amazon Mechanical Turk (AMT).

A dataset of 36232 tweets was introduced as Turkish offensive language by Çöltekin [30]. In this word, the researcher used the approach based on the time interval to retrieve tweets which means that he/she sampled randomly from the Twitter stream for a period of 18 months between April 2018 to September 2019. Tweets in the dataset were annotated similar to Zampieri et al. approach [27] and labeled as offensive or inoffensive at the top level.

Unlike the previous works that used only one social media platform to build a dataset, Jung et al [24] made a new Arabic news comment dataset for offensive language, collected from multiple social media platforms, including Twitter, Facebook, and YouTube. They first collected all news content posted, from 2011 to 2019, by the news agency on their social media accounts. Then, using each content ID, the comments for the content were collected. They obtained manual annotations of these 4000 news comments by the well-known crowdsourcing platform of Amazon Mechanical Turk (AMT) and collected three judgments for each comment. In another word, Romim et al [25] collected 50,281 Bangladesh comments about controversial events that occurred in

Bangladesh following 2017 on Facebook and YouTube platforms. Each comment was annotated by three annotators, and the majority decision was taken as the final decision on offensive or not.

2.2. Persian Offensive Dataset

In this subsection, we review the literature about offensive datasets which is published in the Persian language. POLID is one of the first datasets created by Alavi et al. [21]. They first crawled tweets, Instagram comments, and users' reviews on different Iranian web applications such as Digikala, Snappfood, etc. to build their collection and then annotated text data semi-automatically in two steps. In the first step, they created a basic list of common swearwords and label each text data as 'OFF' (offensive) if it contains at least one element of this list. Otherwise, we categorize it as 'NOT' (Inoffensive). In the second step, they corrected the text data with the wrong label manually. As the result, POLID contains 4,988 text data, comprising 2,453 inoffensive and 2,535 offensive data.

As another Persian dataset for offensive language, one could mention the work done by Mozafari who employed random and lexicon-based sampling to retrieve tweets for a two-month interval from June to August 2020 [22]. In former sampling, tweets were selected randomly and inspected by two experts that revealed that the actual offensive content constituted a maximum of 2% of selected tweets resulting in an unbalanced sampling. In the latter, they benefited from a list of words to filter tweets in order to prevent a bias against some specific topics. The first and second sampling gave them 320k and 200k tweets respectively. They annotate about 6k of the sampled tweets. The label of each tweet was determined similar to the annotation process in [27]. Table I summarizes the reviewed works in terms of the approach of collecting text data and their volume as well as annotation methods and labels.

The main objective of this work is to present a new Persian dataset of offensive language that can be effective in the task of automatic detection of offensive language from social media. One of the most benefits of our dataset is that we employ two approaches based on the users and news-agencies to retrieve a large volume of the comments (near 30k) from Instagram in order to have a wide range of offensive languages in our dataset and it distinguishes our dataset from the existing Persian ones. As another benefit, we can mention that the label of each comment in our dataset was determined by the experts manually in two levels that will be explained in subsection 3.2 latter.

3. Corpus Development

In this section, we detail how the PerBOLD dataset was collected and annotated.

3.1. Data Acquisition

We considered two approaches to collect textual data from the Instagram platform: 1) user-based approach and 2) news agencies-based approach. In the former approach, based on the assumption that the comments published on the controversial users' pages are more likely to be a useful source for offensive language, we first selected some celebrities and political users and then crawled the comments. In the latter, we focused on the comments published in the pages of the news agencies. The reason for this choice was that news content posted on the Instagram platforms attracts interaction between the users and the comments which are the product of users' opinions and beliefs are a strong source for offensive language[24].

The main challenge that we had in collecting data from the Instagram platform was that Instagram imposes many restrictions on robots that extract and crawl data from pages. In order to address this challenge, we benefited from picuki[1] site which has already collected data related to Instagram pages. Unfortunately, this site only crawls the last 24 comments of each post. So pages with a significant number of comments were crawled on Instagram and the rest were crawled on picuki. In the case of posts that were crawled directly from Instagram, up to 612 last comments of each post were collected. The statistical information about collected textual data from Instagram is reported in

Table **II.**

3.2. Annotation Process

The task of automatic detection of offensive speech is usually considered as a supervised classification problem, in which the system is trained on some annotated posts with respect to the presence/absence of some form of offensive content.

Keep in mind that the perception of an offense is subjective, and people may have different opinions on whether the same comment is offensive. In addition, offensive and hateful speech may be used ironically, which obscures the true intent of the author and further confuses taggers. However, the correct understanding of the concept of offensiveness and its lack of eclecticism with other SUD concepts is also very important. Therefore, correct tagging and careful checking of tags are very important and one of the bottlenecks of supplying better performance for machine learning classifiers.

[1] picuki.com

Table I. Comparison of offensive language dataset from social media comments**.**

Reference	Year	Source	Lang.	Method of collection					Volume	Label	Labeling			Number of annotators
				Search by words or hashtags	Time Interval	Search by users	Specific topic	Others			Manual	Automate	Semi automate	
[27]	2019	Twitter	English	*					13241 and 860 training and test samples respectively	Three-level hierarchy annotation scheme 1: Offensive Language Detection (Non-offensive, Offensive) 2: Categorization of Offensive Language (Targeted Insult, Untargeted) 3: Offensive Language Target Identification (Individual, Group, Other)				*
[26]	2020	Twitter	Arabic	*		*		*	5361 tweets	Three-level hierarchy annotation scheme 1: Offensive Language Detection (Non-offensive, Offensive) 2: Hate Language (Hateful, Offensive-Not-Hateful) 3: Categorization of Hateful Language (Religion, Ethnicity, Nationality, Gender)				3
[28]	2020	Twitter	Mexican Spanish	*					10500 tweets	Offensive, Non-offensive	*			
[23]	2020	YouTube	Indian					*	7671 comments	Non-offensive, Offensive (Untargeted, Targeted Individual, Targeted Group, Targeted Other) Non in Kannada language				3 or 5
[30]	2020	Twitter	Turkish		*				36232 tweets	Three-level hierarchy annotation scheme 1: Offensive Language Detection (Non-offensive, Offensive) 2: Categorization of Offensive Language (Targeted Insult, Untargeted) 3: Offensive Language Target Identification (Individual, Group, Other)				2
[24]	2020	Twitter, Facebook, YouTube	Arabic		*		*		4000 news comments	Offensive, Non-offensive				3
[21]	2021	Twitter, Instagram,	Persian					*	4988 text data	Offensive, Non offensive			*	

		some Iranian web applications												
[22]	2021	Twitter	Persian	*	*			*	6k tweets	Three-level hierarchy annotation scheme 1: Offensive Language Detection (Non-offensive, Offensive) 2: Categorization of Offensive Language (Targeted Insult, Untargeted) 3: Offensive Language Target Identification (Individual, Group, Other)				
[25]	2022	Facebook, YouTube	Bangladesh		*		*		50281 comments	Offensive, Non-offensive				3
[29]	2022	Reddit	English				*		6000 comments	Offensive, Non-offensive				4
Our dataset	2022	Instagram	Persian			*	*		28164 comments	Two-level hierarchy annotation scheme 1: Offensive Language Detection (Not Offensive, Offensive) 2: Sexist, Origin, Racist, National, Religion, Political, Others, Sexual, Curse, Degrading	*			3

Since our bottleneck for creating the dataset is manual tagging by humans, we could not label all 151646 comments. Thus, a sample of about 30k comments was randomly selected. These comments were tagged by three taggers. Firstly, two taggers tagged the data. Then, the third tagger, as a linguist expert, judge about inconsistent tagged data. To have a common understanding on offensive speech, regular meetings have been held between the linguist and the two other taggers.

During the tagging procedure, the taggers were asked about a specified comment whether it was an offensive/Non-offensive/advertisement comment. In case of an offensive answer, the annotators were also asked to categorize the offense as curse, insult, sexist, origin, racist, national, religion, political, sexual, and others.

In

Table III, we will see examples of some of the most important tag categories. In the next section, we will examine the statistical information of each of these tags.

Table II. The number of posts and comments crawled from Instagram pages.

Instagram pages	The number of posts	Total number of comments
Fars news agency page on picuki[۲]	696	15018
YJC news agency page on picuki[۳]	612	13852
Iran International page on picuki[۴]	394	9007
Urgent news agency page on picuki[۵]	625	14638
Kara news agency page on picuki[۶]	741	16565
Tn_siasat page on picuki[۷]	1422	10616
Interpreter social media page on picuki[۸]	1260	6105
Superstitions in religion page on picuk[۹]	324	1717
Political users	2698	33602
Celebrities users	580	30526

Table III. Examples of some of the most important tag categories.

Category	Examples
Non-Offensive	چقدر جذابه! همیشه بهترینی اسطوره جان (How at-tractive! Always the best, Dear Myth)
Advertisement	خدمات پرستاری #کاخ_سلامت۷۲۴، ۷ روز هفته در ۲۴ ساعت شبانه روز آماده ارائه خدمات به شما همشهریان عزیز می باشد. (The nursing services of #Kakh_Salamat724 are ready to provide services to you, dear fellow citizens, 24 hours a day, and 7 days a week.)
Offensive-curse words	چقد اوسکلی تی وی پلاس که نشستی کامنت پاک میکنی! (How stupid you are TV Plus, sitting down and deleting people's comments!)
Offensive-personal insult	چرا مردمو مسخره مبکنید ، تو استودیو ضبط میکنید بعد میگی اجرای گروهی ؟ احمقها (Why are you making fun of people, you record in a studio and then you say it's a group performance? Fools)
Offensive-racist	پس تکلیف ۱۵۰ میلیارد دلار که بعداز امضای برجام گرفتید چی شد به مردم که چیزی نرسید شاید عربها ی مفت خور گرفته باشن (So, what happened to the 150 billion dollars that you received after signing the agreement? Nothing was delivered to the people, maybe the sponger Arabs got it)

[2] https://www.picuki.com/profile/fars_news
[3] https://www.picuki.com/profile/yjc.new
[4] https://www.picuki.com/profile/iranintltv
[5] https://www.instagram.com/khabar.fouri
[6] https://www.picuki.com/profile/karanews
[7]https://www.picuki.com/search/tn_siasat
[8] https://www.picuki.com/profile/tarjomaan
[9] www.picuki.com/profile/khorafaat.dar.iran

4. Annotation results and quality

As mentioned before, the total number of gathered comments is 28171. These comments are tagged as non-offensive, offensive, and advertisements. The distribution of tags among the comments is depicted in Fig. 1 and the exact numbers of comments that belong to each of these tags are reported in Table IV.

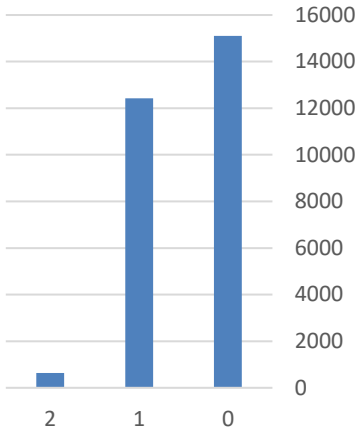

Fig. 1. Distribution of the three tags among comments.

Table IV. The number of comments that belong to each tag.

Tag number	Tag name	Number
0	Non-offensive	15101
1	Offensive	12431
2	Advertisement	632

After a mild pre-processing step, including normalization, number, URL, email address, and special character removal, the distribution of the length of the comments (the number of tokens), not considering the emojis is as shown in Fig. 2. We have not included the lengths with frequencies less than 40 for the purpose of better resolution.

As one can see the lengths are skewed to the shorter values. The most frequent lengths are three and four with the frequencies 2170 and 2147 respectively which follows the common perception of the comments in a social network. This is common to the comments on social networks. Statistics of the comment lengths for the major classes are reported in Table V. As one can see on this table, the length distribution of offensive comments is very similar to the other comments and thus the offensive comments are not distinguishable by just their lengths.

For the offensive comments in the dataset, i.e. the ones with label one, several sub-categories are arranged. The list of these sub-categories is shown in Table VI. The histogram of frequencies is depicted in Fig. 3. The least frequency belongs to the sub-category sexist. The comments with offensive language against a specific race or ethnicity, i.e. origin and racist comment, possess the second smallest frequency. These comments cover about one percent of the gathered comments.

Table V. The length statistics for offensive comments versus the others.

Statistic	Offensive	Others
Average	7.36	7.84
Standard deviation	20.22	19.76
Max	469	458
Min	1	1

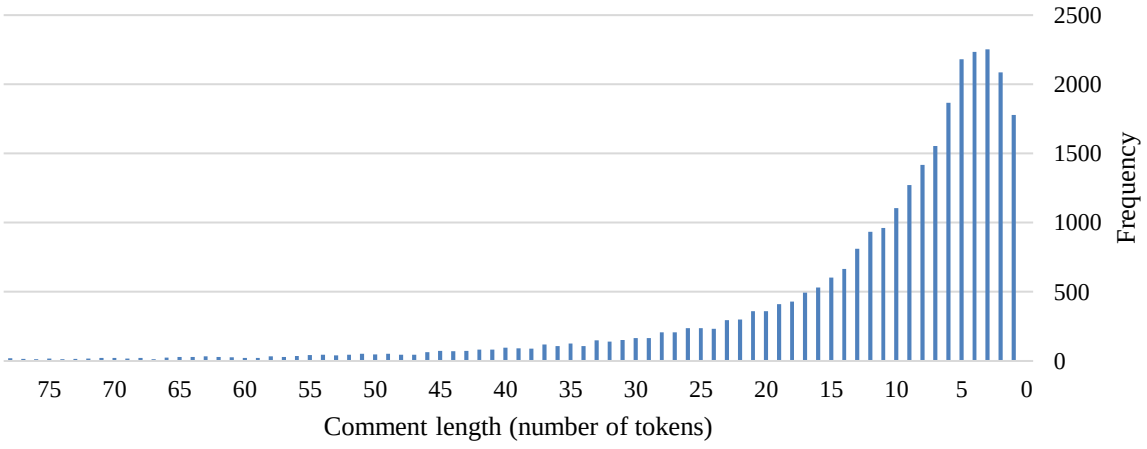

Fig. 2. The distribution comments lengths.

Table VI. The length statistics for offensive comments versus the others.

Name	Description	Number	Percent
Sexist	Offensive speech against a specific gender	20	0.16
Origin	Offensive speech against a specific place of origin	43	0.35
Racist	Offensive speech against a specific ethnicity	83	0.67
National	Offensive speech against national beliefs and achievements	137	1.01
Religion	Offensive speech against religious beliefs	236	1.90
Political	Offensive speech against a political party	607	4.88
Others	Any other kind of offensive speech such as threats or slander	759	6.11
Sexual	Offensive speech containing sexual talks or insults	1286	10.351
Curse	Offensive speech containing curse words or	3378	27.13
Degrading	Offensive speech for humiliating someone	5887	47.36

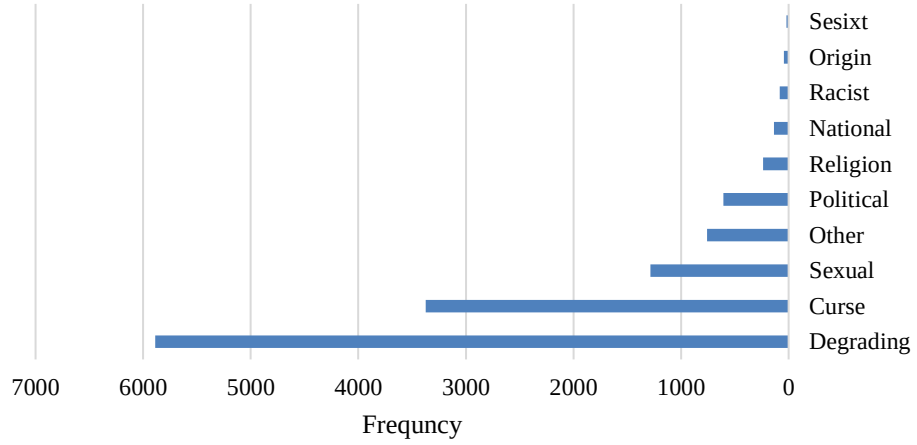

Fig. 3. Histogram of frequencies for offensive sub-categories.

In order to be accompanied, we apply two variations of naïve Bayes (e.g. Bernoulli NB, and Gaussian NB), as well as logistic regression, as some baseline text categorization models to demonstrate the coherence and effectiveness of the proposed dataset. In these models we use TF-IDF method to construct the feature vectors. We remove 10% of data (about 3000 data) to be used as test data. Table VII shows the results of applying these methods on test data.

Table VII. Results of Applying some Baseline Categorization Models on the Dataset.

Method\criteria	Precision	Recall	F1-score	Accuracy
Bernoulli NB	0.76	0.75	0.75	0.76
Gaussian NB	0.74	0.73	0.73	0.73
Logistic Regression	0.74	0.73	0.73	0.73

5. Conclusion

In this paper, we provide PerBOLD, a dataset of the Instagram comments in Persian which is useful to detect offensive language automatically from the text. We followed a two-level process for annotating the comments in the dataset manually. In the first level, the tag of each comment is determined based on whether it has offensive language words or not. In the second level, the offensive comments are categorized into ten categories based on their intent. To the best of our knowledge, PerBOLD is the first Persian dataset with a huge number of comments and annotated with very fine-grained tags of different types of offensive. In order to be accompanied, we apply baseline text categorization models (naïve bayes, and logistic regression) to demonstrate the coherence and effectiveness of the proposed dataset. However, as future work, state-of-the-art text classification methods, as well as word embedding methods, could be used to better identify offensive data utilizing from the provided dataset as a training dataset.

6. References

[1] F. Ghanbari, M. Rahmani, "Presenting a Semantic Orientation Based Method for Multi-Label Classification of Movies Content Using Their Subtitle Texts", *Tabriz Journal of Electrical Engineering*, vol. 47, pp. 1599-1611, 2018.

[2] Z. Amighi, M. Yousef Sanati, M. Dezfoulian, "DynamicEvoStream: An EvoStream based Algorithm for Dynamically Determining The Number of Clusters in Data Streams", *Tabriz Journal of Electrical Engineering*, vol. 51, pp. 315-326, 2022.

[3] H. Mulki, H. Haddad, C. B. Ali, H. Alshabani, "L-hsab: A levantine twitter dataset for hate speech and abusive language", in *Proceedings of*

the third workshop on abusive language online, Florence, Italy, pp. 111-118, 2019.

[4] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, "Abusive language detection in online user content", in *Proceedings of the 25th international conference on world wide web*, Montréal Québec Canada, pp. 145-153, 2016.

[5] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior", in *Twelfth International AAAI Conference on Web and Social Media*, Palo Alto, California, USA, 2018.

[6] P. Liu, J. Guberman, L. Hemphill, A. Culotta, "Forecasting the presence and intensity of hostility on Instagram using linguistic and social features", in *Twelfth international aaai conference on web and social media*, Palo Alto, California, USA, 2018.

[7] R. Sprugnoli, S. Menini, S. Tonelli, F. Oncini, E. Piras, "Creating a whatsapp dataset to study pre-teen cyberbullying", in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium, pp. 51-59, 2018.

[8] H. Zhong, H. Li, A. C. Squicciarini, S. M. Rajtmajer, C. Griffin, D. J. Miller, C. Caragea, "Content-Driven Detection of Cyberbullying on the Instagram Social Network", in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 3952-3958, 2016.

[9] J. Qian, A. Bethke, Y. Liu, E. Belding, W. Y. Wang, "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*, Hong Kong, China, pp. 4755–4764, 2019.

[10] O. De Gibert, N. Perez, A. García-Pablos, M. Cuadros, "Hate speech dataset from a white supremacy forum", in *Proceedings of the 2nd Workshop on Abusive Language Online ({ALW}2)*, Brussels, Belgium, pp. 11--20, 2018.

[11] Ò. G. i Orts, "Multilingual detection of hate speech against immigrants women in Twitter at SemEval-2019 task 5: Frequency analysis interpolation for hate in speech detection", in *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 460-463, 2019.

[12] M. A. Bashar, R. Nayak, K. Luong, T. Balasubramaniam, "Progressive domain adaptation for detecting hate speech on social media with small training set and its application to COVID-19 concerned posts", *Social Network Analysis and Mining*, vol. 11, pp. 1-18, 2021.

[13] X. Huang, L. Xing, F. Dernoncourt, M. J. Paul, "Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition", in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, pp. 1440--1448, 2020.

[14] P. Fortuna, J. R. da Silva, L. Wanner, S. Nunes, "A hierarchically-labeled portuguese hate speech dataset", in *Proceedings of the third workshop on abusive language online*, pp. 94-104, 2019.

[15] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval) ", in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA, pp. 75--86, 2019.

[16] G. Kennedy, A. McCollough, E. Dixon, A. Bastidas, J. Ryan, C. Loo, J. Ryan, C. Loo, S. Sahay, "Technology solutions to combat online harassment", in *Proceedings of the first workshop on abusive language online*, pp. 73-77, 2017.

[17] M. Wiegand, M. Siegel, J. Ruppenhofer, "Overview of the germeval 2018 shared task on the identification of offensive language", in *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing*, Vienna, Austria, 2018.

[18] I. Markov, N. Ljubešić, D. Fišer, W. Daelemans, "Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection", in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 149-159, 2021.

[19] F. Alves Vargas, I. Carvalho, F. Rodrigues de Góes, F. Benevenuto, T. Alexandre Salgueiro Pardo, "Building an Expert Annotated Corpus of Brazilian Instagram Comments for Hate Speech and Offensive Language Detection", *arXiv e-prints*, p. arXiv: 2103.14972, 2021.

[20] M. Wiegand, M. Siegel, J. Ruppenhofer, "Overview of the germeval 2018 shared task on the identification of offensive language", in *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, pp. 1-10, 2018.

[21] P. Alavi, P. Nikvand, M. Shamsfard, "Offensive Language Detection with BERT-based models, By Customizing Attention Probabilities", *arXiv preprint arXiv:2110.05133*, 2021.

[22] M. Mozafari, "Hate speech and offensive language detection using transfer learning approaches", Institut Polytechnique de Paris, 2021.

[23] A. Hande, R. Priyadharshini, B. R. Chakravarthi, "KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection", in *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pp. 54-63, 2020.

[24] S. A. Chowdhury, H. Mubarak, A. Abdelali, S.-g. Jung, B. J. Jansen, J. Salminen, "A multi-platform Arabic news comment dataset for offensive language detection", in *Proceedings of*

the 12th language resources and evaluation conference, pp. 6203-6212, 2020.

[25] N. Romim, M. Ahmed, M. Islam, A. S. Sharma, H. Talukder, M. R. Amin, "BD-SHS: A Benchmark Dataset for Learning to Detect Online Bangla Hate Speech in Different Social Contexts", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, pp. 4755--4764, 2022.

[26] S. Alsafari, S. Sadaoui, M. Mouhoub, "Hate and offensive speech detection on Arabic social media," *Online Social Networks and Media*, vol. 19, p. 100096, 2020.

[27] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, "Predicting the type and target of offensive posts in social media," in *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologie*, Minneapolis, Minnesota, pp. 1415--1420, 2019.

[28] M. J. Díaz-Torres, P. A. Morán-Méndez, L. Villasenor-Pineda, M. Montes, J. Aguilera, L. Meneses-Lerín, "Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset", in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 132-136, 2020.

[29] R. Hada, S. Sudhir, P. Mishra, H. Yannakoudakis, S. M. Mohammad, E. Shutova, "Ruddit: Norms of offensiveness for English Reddit comments", in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 2700--2717, 2021.

[30] Ç. Çöltekin, "A corpus of Turkish offensive language on social media", in *Proceedings of the 12th language resources and evaluation conference*, pp. 6174-6184, 2020.