Hindawi
Complexity
Volume 2020, Article ID 6684995, 7 pages
https://doi.org/10.1155/2020/6684995

Research Article

USAD: An Intelligent System for Slang and Abusive Text Detection in PERSO-Arabic-Scripted Urdu

Nauman Ul Haq,[1] Mohib Ullah,[1] Rafiullah Khan ,[1] Arshad Ahmad ,[2] Ahmad Almogren,[3] Bashir Hayat,[4] and Bushra Shafi[5]

[1]*Intitute of Computer Science and Information Technology, The University of Agriculture, Peshawar 25000, Pakistan*
[2]*Department of IT & Computer Science, Pak-Austria Fachhochschule: Institute of Applied Sciences & Technology, Mang Khanpur Road, Haripur 22620, Pakistan*
[3]*Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11633, Saudi Arabia*
[4]*Institute of Management Sciences, Peshawar 25000, Pakistan*
[5]*Department of Rural Sociology, The University of Agriculture, Peshawar 25000, Pakistan*

Correspondence should be addressed to Arshad Ahmad; yaarshad@gmail.com

Received 9 October 2020; Revised 19 November 2020; Accepted 23 November 2020; Published 30 November 2020

Academic Editor: M. Irfan Uddin

The use of slang, abusive, and offensive language has become common practice on social media. Even though social media companies have censorship polices for slang, abusive, vulgar, and offensive language, due to limited resources and research in the automatic detection of abusive language mechanisms other than English, this condemnable act is still practiced. This study proposes USAD (Urdu Slang and Abusive words Detection), a lexicon-based intelligent framework to detect abusive and slang words in Perso-Arabic-scripted Urdu Tweets. Furthermore, due to the nonavailability of the standard dataset, we also design and annotate a dataset of abusive, offensive, and slang word Perso-Arabic-scripted Urdu as our second significant contribution for future research. The results show that our proposed USAD model can identify 72.6% correctly as abusive or nonabusive Tweet. Additionally, we have also identified some key factors that can help the researchers improve their abusive language detection models.

1. Introduction

The birth of social media entirely revolutionized the ways and purpose of mass communication [1]. In the early days, mass communication media was used with the ethical and moral responsibilities as governed by social norms. Besides that, mass communication media was effectively used for education and training. Social media currently allows every connected individual to express their feelings about anything using Twitter, Facebook, Instagram, blogs, or other social media sites [2]. Recent studies about social media show peoples' lack tolerance that turns into aggression through which people use such language that may offend others' feelings [3]. However, most of the social media sites have policies for content publications and penalties for policy violations.

Nonetheless, in the case of informal and vulgar textual content, the policies' violation is hard to detect manually [1] due to an immense number of posts. Moreover, these social media sites also allow users to post their textual content in native languages. According to the survey conducted in 2018, the English language is used in only 32% of all Tweets [4]. Twitter is a microblogging and social networking service that allows users to express their views using a Tweet of 280 characters [2].

Hate speech and offensive and abusive language detection on social media is an active research field [5]. There are several studies available for hate speech abusive language detection for English [6, 7], Danish [8], Arabic [1, 9], Indonesian [10], and others. These studies used different methodologies to detect abusive and offensive language, such as lexicon-based detection in the English language [11], n-gram for English [12] and Roman Urdu [5] language, pattern matching [13], blacklist [7], and others. However, to the best of our knowledge, abusive, offensive, and slang word detection from the Urdu language in Perso-Arabic script has not been performed due to its complexity.

The Urdu language is challenging due to its morphological and syntactical complexity as the Urdu language draws grammatical structures and vocabulary from Persian, Sanskrit, Arabic, and Turkish [14]. Due to morphological and syntactical complexity, minimal research has been conducetd on Urdu text, especially for abusive word detection [5]. Similarly, there is no standard dataset in Perso-Arabic-scripted Urdu publicly available for offensive text detection.

This study proposes a lexicon-based framework to detect abusive, offensive, and slang words in Perso-Arabic-scripted Urdu Tweets. Additionally, due to the nonavailability of the standard dataset, we also design and annotate a dataset composed of abusive, offensive, and slang word Perso-Arabic-scripted Urdu as our second significant contribution for future research. The results show that our proposed USAD model can identify 72.6% of Tweets as abusive or nonabusive correctly with the precision of 55.21%. The contributions of this work are as follows:

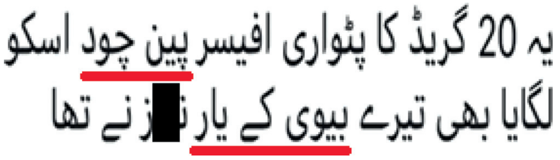
(1) A lexicon-based framework that detects abusive, offensive, and slang words in Perso-Arabic-scripted Urdu Tweets is proposed

(2) A dataset composed of abusive, offensive, and slang words Perso-Arabic-scripted Urdu is designed and annotated

In Section 2, a brief introduction of the Urdu language is given. Related work is discussed in Section 3, while in Section 4, we discussed the USAD model. The experimentation preliminaries are discussed in Section 5, while in Section 6 and 7, we discussed the results and conclusions with future recommendations, respectively.

1.1. Motivation.

Urdu is one of the main spoken languages in the subcontinent and the 11th most spoken language in the world [15]. Urdu is also the national and official language of Pakistan, and most of the users in Pakistan use Perso-Arabic-scripted Urdu on social media. Like other countries, users of social media in Pakistan often use slang and vulgar words in their Tweets [5]. As shown in Figure 1, a single Tweet in the Urdu language contains two offensive words. According to the PECA '16 (Prevention of Electronic Crimes Act, 2016) Pakistan chapter ii, section 20 and 21, offence against dignity and modesty of a person using any communication medium is a punishable act [16]. Unfortunately, no mechanism is available to automatically detect the Urdu language's offensive and vulgar words in the Perso-Arabic script. Even though the use of abusive, offensive, and vulgar language in Tweets is punishable according to Pakistan's laws, this condemnable act is still in practice due to the nonavailability of automatic detection mechanisms.



یہ 20 گریڈ کا پٹواری افیسر پین چود اسکو
لگایا بھی تیرے بیوی کے یار[redacted]ز نے تھا

Figure 1: Tweet in Perso-Arabic-scripted Urdu with abusive words.

2. Urdu and Perso-Arabic Script

Urdu is one of the South Asian region's popular languages and Pakistan's national and official language [15]. Urdu belongs to the Indo-Aryan language family, and colloquially, it is mostly mutually intelligible with conversational Hindi [17]. Formal Urdu draws grammatical structures and vocabulary mainly from the Persian language and a small amount of Sanskrit, Arabic, and Turkish language [14]. Like Persian and Arabic, Urdu is written from right to left in Perso-Arabic script and Urdu has more phonic sounds than Arabic and Persian. Urdu has 40 distinct alphabets called "Huruf-e-Tahaji," written in various calligraphic styles such as Nastaliq, Naskh, Reqa, Diwani, and others [18].

Similarly, Hindi, a mutually intelligible language of the Urdu language, is written in Devanagari script [19]. Before the development of the Urdu charter set and keyboard, the Roman script was used to write the contents in Urdu. Urdu written in Roman script is called Roman Urdu [20] and Romanagari for Hindi [5].

Due to complex morphological and grammatical structures, diacritics [21], and limited linguistics resources, the Urdu language is mostly neglected by the research community. In this regard, the first ever 8 bit encoding standard for Urdu, "Urdu Zabta Takhti (UZT) 1.01," was developed and accepted by the Government of Pakistan in 2000 [22]. Several studies are available on the Urdu language, such as opinion mining, sentiment analysis, text clustering, and classification. In contrast, only a single study is available for offensive language detection in Roman Urdu [5]. Therefore, the detection of abusive and offensive language in Perso-Arabic-scripted Urdu is still an open issue.

3. Related Work

An increasing amount of attention to the computational linguistic community has been given to the automatic detection of hate speech, slang words, and offensive and abusive language from online social media. Social media is an open forum where people from different countries, races, nationalities, religions, and cultures can share their opinions and comments. These comments might usually include offensive or abusive words against other users [11]. Therefore, it is a crucial issue to detect and block or censor this condemnable practice. In this regard, many studies have been conducted in English [6, 7], Arabic [1, 9], Indonesian [10], and Roman Urdu [5]. This section briefly discusses various studies conducted for automatic hate speech and offensive language detection on social media for different languages.

Recently, automatic detection of hate speech and offensive and abusive words in users' comments has become a trending research topic. Researchers have used different methodologies such as machine learning techniques, lexicon-based techniques, graph-based techniques, and others to detect abusive words automatically.

Watanabe et al. used unigram and patterns with supervised learning algorithms to classify hateful and clean comments in English [6]. Burnap and Williams used supervised machine learning and a statistical model to detect hate speech on Twitter [23]. They used a combination of rule-based, spatial-based, and probabilistic classifiers to detect the hate speech. Lee et al. proposed a model for abusive word detection using a dictionary of abusive and nonabusive words and unsupervised learning techniques for social media comments in the English language [7].

Chen et al. proposed the Lexical Syntactic Feature (LSF) architectures that use specific bulling content, writing style, and structure as a feature vector to predict the user's potentiality for creating obscene content [24]. Park and Fung used a Character-level Convolutional Network, Word-level Convolutional Network, and Hybrid Convolutional Network to detect racist, sexist, and abusive Tweets in the English language [25], while Mishra et al. used a Graph Convolutional Network with the user's online community structure and linguistic behavior to detect the offensive language [26].

While most approaches work on the English language, some studies are available for other languages; for example, Pelle et al. proposed the "Hate2Vec" approach that uses lexicon and bag-of-word classifiers to detect offensive comments in English and Portuguese languages [27]. Sigurbergsson and Derczynski proposed the Recurrent Neural Network-based hate offensive language and speech detection model for Danish and English languages [8]. In contrast, Schneider et al. used a Convolutional Neural Network to detect the abusive, insulting, and offensive comments for the German Language [28].

Ibrohim and Budi proposed n-gram and supervised learning-based approaches to detect the abusive language in Indonesian social media [10]. Alakrot et al. proposed an n-gram-based model to catch even misspelled offensive and obscene Arabic words and phrases in user comments [9]. For this purpose, they also construct a dataset of abusive words in the Arabic language to detect antisocial behavior [29]. In comparison, Abozinadah proposed a multidimensional analysis model that uses social graph analysis, statistical analysis, and lexical analysis to detect the abusive language in Arabic text [1]. Akhter et al. proposed an n-gram and supervised machine learning algorithm-based offensive language detection model for Roman Urdu [5].

Rizwan et al. proposed a CNN-gram-based model to detect hate speech and offensive language in Roman Urdu [30]. They tested their model on the RUHSOLD (Roman Urdu Hate Speech and Offensive Language Detection) dataset. They tested the performance of their proposed model with and seven baseline models. Abbas performed experiments using multiple machine learning algorithms to detect toxic (offensive) comments in Roman Urdu [31]. He reported that Random Forest gave 96.4% accuracy with the character 4-gram technique. Kausar et al. proposed "Pro-SOUL," a framework to identify the propaganda in online Urdu content [32]. They proposed a Linguistic Inquiry and Word Count Dictionary to detect psycholinguistic features to propaganda in Urdu contents.

Offensive language and hate speech detection is an important issue, especially in social media, which can influence the user's behavior and reaction. Unfortunately, most of the research has focused on automatic offensive language detection for resource-rich languages [5] such as English, Danish, and German, while publication on other languages is rare. Therefore, we proposed automatic abusive and slang word detection for the Perso-Arabic Urdu language in this research.

4. Urdu Slang and Abusive Word Detection (USAD) Model

Urdu is one of the major languages of the subcontinent and the national language of Pakistan. Most of the users of social media from Pakistan prefer to comment in their native Urdu language in Perso-Arabic script. Like other languages, the use of abusive and offensive phrases is very much common in Urdu comments. In this section, we discuss the working of the proposed USAD model.

4.1. Working of the USAD Model. The proposed USAD model is divided into two major phases, i.e., the Lexicon Building Phase and the data testing phase. We crawled and collected Tweets posted in Perso-Arabic script using Twitter Application Programming Interface (API) in the lexicon building phase. For dataset preparation, the tweets are saved in a text file using UTF-8 (Unicode Transformation Format version 8) encoding and forwarded to the preprocessing module. In the data preprocessing step, stop words, punctuation marks, digits, and nonlanguage characters are removed, and tweets are tokenized as a single entity. After data cleansing, a lexicon of Urdu abusive and slang words is created manually. The details of lexicon creation are discussed in Section 5.1. In the data testing phase, clean tweets are given as input to the classification module. In the classification module, each word of an input tweet is tested against the abusive words lexicon for classifying tweets as abusive or nonabusive. The architecture of the proposed USAD model is shown in Figure 2.

5. Experimentation Preliminary

A Python-based lexicon building and the testing tool are developed to implement the proposed USAD model for abusive and offensive Urdu tweet detection. The abusive word dictionary is used to classify the tweets into abusive and nonabusive class. This section explains the methods of Urdu abusive and slang words lexicon and testing dataset

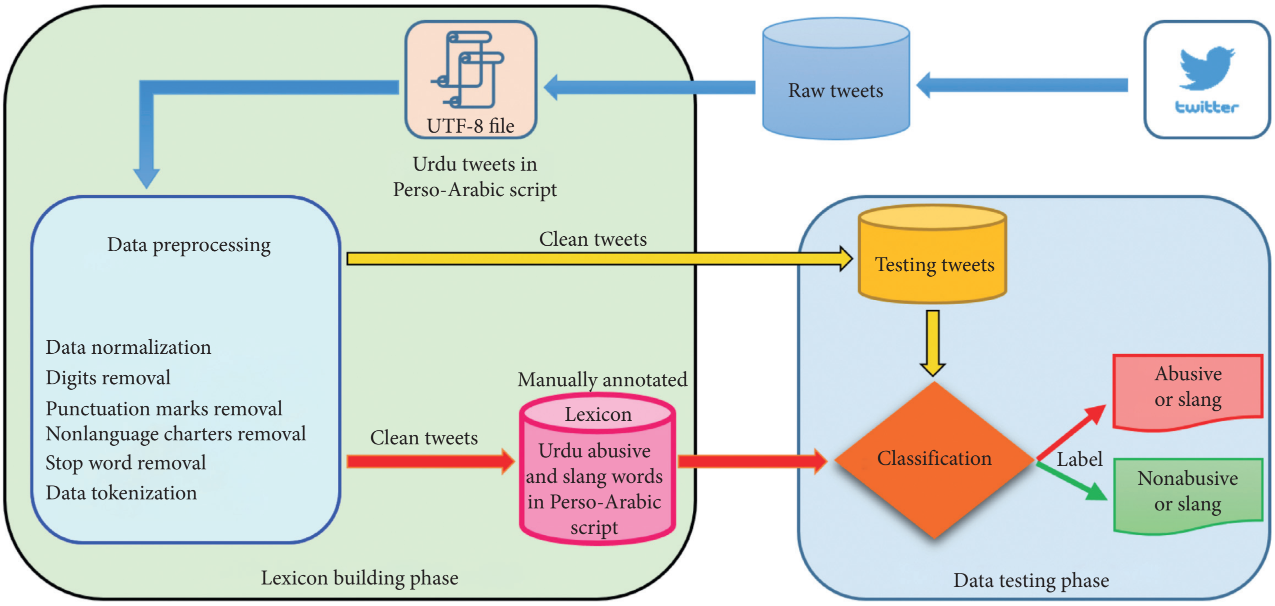


Figure 2: The architecture of the proposed Urdu Slang and Abusive Words Detection (USAD) model.

creation. Furthermore, to evaluate the proposed USAD model's performance, we used a standard machine learning performance evaluation parameter, i.e., precision, recall, F-measure, and accuracy.

5.1. Dataset and Lexicon Creation. For abusive and slang words lexicon, we collected more than 5000 Tweets and replies posted by famous politicians, journalists, analysts, and intellectuals on different topics during October 2019 and December 2019. The tweets are then saved into a text file with a UTF-8 encoding scheme. After applying the preprocessing and data cleansing steps, an abusive words lexicon is created manually. The abusive word lexicon is composed of 2533 abusive words of Urdu language posted in 3 months period.

For testing, we build a dataset composed of 1200 Tweets and replies posted on different topics during the same period, i.e., October 2019 and December 2019. For data cleansing, the same preprocessing steps are used. After data cleansing, we manually annotate the dataset into abusive and nonabusive classes for result comparison. After manual annotation, the testing data are supplied to the classification algorithm that uses a string-matching method for Tweet classification. The details of lexicon building a dataset and testing dataset are given in Table 1. Similarly, the examples of Tweets with abusive words in Perso-Arabic Urdu are given in Table 2.

5.2. Performance Evaluation Metrics. For performance evaluation of the proposed USAD model for detecting abusive and slang Urdu tweets, we used standard machine learning performance evaluation parameters, i.e., precision, recall, F-measure, and accuracy. Precision shows how many of the identified tweets are abusive, and recall shows how many of the total tweets are correctly identified as abusive tweets, while F-measure is a harmonic mean of precision and recall [20,33]. The equations of the selected evaluation parameters are given as follows:

Table 1: Lexicon and testing dataset properties.

Lexicon data	Total no. of tweets	5220
	Total no. of abusive words	1250
	Period	October–December 2019
Testing data	Total no. of tweets	1200
	Abusive tweets	365
	Nonabusive tweets	835
	Period	October–December 2019

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2}$$

$$F - \text{Measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{4}$$

where TP stands for true positive, FP stands for false positive, TN stands for true negative, while FN stands for the false-negative sample.

6. Results and Discussions

This research proposes a lexicon-based framework to detect abusive, offensive, and slang words in Perso-Arabic-scripted Urdu Tweets. For experimentation, we build a Python-based

Table 2: Examples of Tweets with abusive words in Perso-Arabic Urdu.

Tweet	Abusive words
حیثت ہی ہوگئی کہ ٹویٹ صرف 60، 70 لوگ لائیک کرتے ہیں کچھ رذیل تمہیں کوریج نہ دیں تو تیرا نام و نشان نہیں ہوگا	رذیل
لے بھائی احٔنی کی وکالت- مأمد اور عمٔمہ- کنجروں کی وکالت کنجر-	کنجروں ، کنجر
مأمد تیری چوت میں تو خارش شروع ہو گئی ہو گی	چوت
بےآہ کا لیول لوکل نہیں ہے انٹرنیشنل لیول کی گشتی ہے- بالکل اپ کی طرح	گشتی
ایک تو ساتھ فورم والے کنجروں اور کنجریوں کی گانڈ پہ ہر وقت فٔج چڑھی رہتی ہے	کنجروں، کنجریوں، گانڈ
لو میرا جسم میری مرضی والی دو ٹکے کی رنڈئ بھی فٔج کے خلاف بھونکنے آئ ھے	رنڈئ
انشاءاللہ چؤری صاحب ایسے ھی زلیل ھو کر نکلےگا	زلیل

testing environment for both lexicon building and Tweets classification (Section 5). For lexicon building, we crawled more than 5000 Urdu Tweets and made a lexicon of 1250 abusive and slang words, while for testing, we took 1200 Urdu Tweets and manually annotated them into abusive and nonabusive classes. After manual classification, the dataset is supplied to the testing module for the automatic classification of the data using an abusive lexicon. This section discusses the results of the proposed USAD model's effectiveness in the automatic detection of abusive Urdu Tweet.

The results show that, out of 365 abusive Tweets, our proposed USAD model correctly identified 265 Tweets as abusive, while out of 835, the proposed USAD identified 620 Tweets as nonabusive Tweets. The results are actual, and the predicted Tweets are given in Table 3. In terms of precision and recall, USAD performed well by identifying 72.6% (Recall) Tweets as abusive correctly with the precision of 55.21%. Similarly, the proposed USAD model was able to classify 73.75% of Tweets correctly as abusive and nonabusive. The precision, recall, f-measure, and accuracy of the proposed USAD model are depicted in Figure 3.

The USAD model was able to identify 72.6% of Tweets as abusive and 74.3% Tweets as nonabusive correctly. Upon investigation of misclassified Tweets, it was found that the Tweets were misclassified due to the limited abusive words lexicon, proverbs, and quotes, contextual abusive words, abusive terms of other languages, and misspelled abusive words. The details of the findings mentioned above are discussed in this section with examples in Table 4.

6.1. *Limited Abusive Words Lexicon.*

One of the significant limitations of Tweets' misclassification is the number of abusive words in the abusive words lexicon. The creation of new slang, abusive, and vulgar terms is an ongoing process in any language, just like literature. Additionally, some slang words are event-driven, and some are contextual. For this research, we took 5220 Tweets in the Urdu language to develop an abusive language lexicon composed of 1250 abusive words posted in three months (October to December 2019).

6.2. *Proverbs and Quotes.*

Another major issue of misclassification is the lack of proverbs and quotes in the abusive words' lexicon. Urdu has a rich history stemming from diverse cultures coexisted for a long period within the same region. As Urdu is also the combination of various languages of the subcontinent, it has a rich collection of proverbs and quotes of almost every language, especially Persian. Some of the proverbs and quotes contain abusive words which actually are nonabusive words. As in this research, we build our lexicon for abusive words; therefore, many positives quotes and proverbs are identified as abusive Tweets.

Table 3: Confusion matrix.

		Actual	
		Abusive	Nonabusive
Predicted	Abusive	265	215
	Nonabusive	100	620

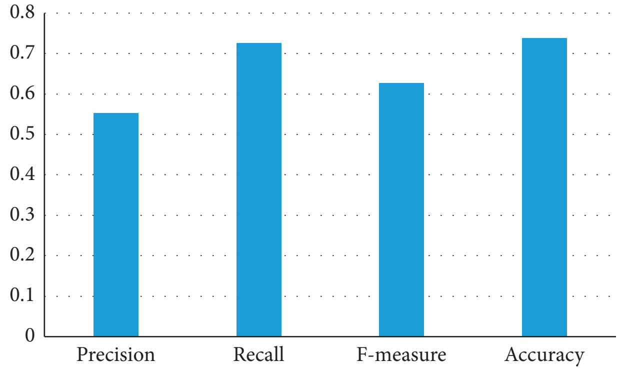


Figure 3: Precision, recall, F-measure, and accuracy of the proposed USAD model.

6.3. *Contextual Abusive Words.*

Most of the languages' abusive words usually evolve from an event or context, such as sarcasm. People typically make fun of the actions or statements of an individual to defame them. These words are contextual abusive words which are hard to identify due to a lack of contextual information.

6.4. *Abusive Terms of Other Languages.*

Another important reason for data misclassification is abusive words other than Urdu or in Roman script. Most users also use abusive words in their native languages such as Pashto, Punjabi, Sindhi, and Hindko or use abusive words of the English language in their Tweets. As our Lexicon is Urdu based, Tweets with abusive words in other languages are identified as nonabusive.

Table 4: Confusing Tweets with their types.

Tweet	Actual	Identified as	Type
وہی اللہ ہے جس کو چاہے عزت دے جس کو چاہے زلیل کرے	Nonabusive	Abusive	Quote
دراصل بھینسوں کا چارہ سستا اورکتوں کےلئے گوشت مہنگا آتا ہے	Nonabusive	Abusive	Contextual sarcastic
نوۓ اور بلؔ حجڑہ بھی بہت بکواس بک رہی تھے-	Abusive	Nonabusive	Contextual sarcastic
مرڈؔ بڑا کھوتی دا پترھے	Abusive	Nonabusive	Panjabi language
کہوتہ کہیں کا	Abusive	Nonabusive	Misspell
ہی fucking چورسب کچھ لوٹ کرلے گئے	Abusive	Nonabusive	Roman-scripted English
دھوبی کا کتا نہ گھر کا نہ گھاٹ کا	Nonabusive	Abusive	Proverb

6.5. Misspell Abusive Word. Currently, the misspelling of words is a prevalent practice on social media. Users are usually careless about their posts as the human brain can process and infer the text's meaning. Therefore, it is difficult for the computer to identify abusive words with wrong spelling.

7. Conclusions

In this work, we proposed the USAD model for automatic detection of abusive Tweets posted in Perso-Arabic-scripted Urdu. For experimentation, we used a lexicon of abusive Urdu words composed of 1250 words and a testing dataset consisting of 1200 manually annotated Tweets (365 abusive and 835 nonabusive). The results show that the proposed USAD model can identify 72.6% of Tweets as abusive or nonabusive correctly with the precision of 55.21%. Upon the investigation of misclassified Tweets, we have found that the Tweets were misclassified due to the limited abusive words lexicon, nonexistence of Urdu proverbs and quotes in the lexicon, contextual abusive word, abusive terms of other languages, and misspelled abusive words.

It is concluded that the proposed USAD model's performance can be improved with the more significant abusive lexicon. Moreover, the inclusion of the abusive terms of other languages of Pakistan such as Pashto, Punjabi, English, and others can also improve the proposed model's performance as people usually prefer abusive terms of their mother tongue. Similarly, the inclusion of all possible misspelled abusive words in a lexicon will also significantly improve the model's performance. Furthermore, a lexicon of proverbs and quotes and an abusive lexicon can help the model decide the class of the phrase. However, for this purpose, a phrase-level matching will be appropriate. The most challenging task will be handling contextual abusive terms and sarcastic terms as these terms are based on some event, and usually, the words used in these Tweets replies are not directly connected with the base Tweets.

For future work, we aimed to enhance the lexicon of abusive words with more abusive words, including abusive words from other local languages in both Perso-Arabic and Roman scripts. We have also aimed to create a lexicon for proverbs and quotes of the Urdu language to improve the machine's performance by differentiating abusive words and quotes or proverbs. To solve the problem of misspelling abusive words, edit distance, and *n*-gram approaches are potential candidates. For solving the sarcastic and contextual abusive words problem, the semantic graph-based method can be used effectively.

Data Availability

All the data used to support the findings of the study are available in the manuscript.

Conflicts of Interest

All the authors declare that they have no conflicts of interest related to this study.

Acknowledgments

The authors would like to thank and acknowledge the support provided by King Saud University, Riyadh, Saudi Arabia, through Researchers Supporting Project number RSP-2020/184. This work was supported by King Saud University, Riyadh, Saudi Arabia, through Researchers Supporting Project number RSP-2020/184.

References

[1] E. Abozinadah, *Detecting Abusive Arabic Language Twitter Accounts Using a Multidimensional Analysis Model*, George Mason University, Fairfax, VA, USA, 2017.

[2] H. Mubarak and K. Darwish, "Arabic offensive language classification on twitter," in *Proceedings of the International Conference on Social Informatics*, pp. 269–276, National Research Council of Pisa, Pisa, Italy, May 2019.

[3] K. Stapleton, "Swearing and perceptions of the speaker: a discursive approach," *Journal of Pragmatics*, vol. 170, pp. 381–395, 2020.

[4] "2018 research on 100 million tweets: what it means for your social media strategy for twitter," *Vicinitas.*, https://www.vicinitas.io/blog/twitter-social-media-strategy-2018-research-100-million-tweets#:%7E:text=Tens%20of%20languages%20are%20used,used%20in%20posting%20Twitter%20messages, 2018.

[5] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, and M. T. Sadiq, "Automatic detection of offensive language for Urdu and roman Urdu," *IEEE Access*, vol. 8, pp. 91213–91226, 2020.

[6] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on twitter: a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE Access*, vol. 6, pp. 13825–13835, 2018.

[7] H.-S. Lee, H.-R. Lee, J.-U. Park, and Y.-S. Han, "An abusive text detection system based on enhanced abusive and non-

abusive word lists," *Decision Support Systems*, vol. 113, pp. 22–31, 2018.

[8] G. I. Sigurbergsson and L. Derczynski, "Offensive language and hate speech detection for Danish," 2019, https://arxiv.org/abs/1908.04531.

[9] A. Alakrot, L. Murray, and N. S. Nikolov, "Towards accurate detection of offensive language in online communication in Arabic," *Procedia Computer Science*, vol. 142, pp. 315–320, 2018.

[10] M. O. Ibrohim and I. Budi, "A dataset and preliminaries study for abusive language detection in Indonesian social media," *Procedia Computer Science*, vol. 135, pp. 222–229, 2018.

[11] N. D. Gitari, Z. Zhang, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 4, pp. 215–230, 2015.

[12] P. Rani and A. K. Ojha, "KMI-coling at SemEval-2019 task 6: exploring *N*-grams for offensive language detection," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 668–671, Minneapolis, MN, USA, June 2019.

[13] A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, "Offensive language detection using multi-level classification," in *Proceedings of the Canadian Conference on Artificial Intelligence*, pp. 16–27, May Ottawa, ON, Canada, May 2010.

[14] T. Kiss and A. Alexiadou, *Syntax-Theory and Analysis*, Walter de Gruyter GmbH & Co KG, Berlin, Germany, 2015.

[15] M. Sharjeel, R. M. A. Nawab, and P. Rayson, "COUNTER: corpus of Urdu news text reuse," *Language Resources and Evaluation*, vol. 51, no. 3, pp. 777–803, 2017.

[16] "Prevention of electronic Crimes act, ," Pakistan N. A. SECRETARIAT, Karachi, Pakistan, 2016.

[17] C. Everaert, *Tracing the Boundaries Between Hindi and Urdu: Lost and Added in Translation Between 20th Century Short Stories*, Brill, Leiden, Netherlands, 2010.

[18] J. Bloom, S. S. Blair, and S. Blair, *Grove Encyclopedia of Islamic Art & Architecture: Three-Volume Set*, Oxford University Press on Demand, Oxford, UK, 2009.

[19] C. P. Masica, *The Indo-Aryan Languages*, Cambridge University Press, Cambridge, UK, 1993.

[20] M. Daud, R. Khan, and A. Daud, *Roman Urdu Opinion Mining System (RUOMiS)*, https://arxiv.org/abs/1501.01386, 2015.

[21] S. Hussain, S. Gul, and A. Waseem, "Urdu encoding and collation sequence for localization,"Center for Research in Urdu Language Processing National University of Computer and Emerging Sciences, Lahore, Pakistan.

[22] S. Hussain and M. Afzal, "Urdu computing standards: Urdu zabta takhti (uzt) 1.01," in *Proceedings of the IEEE International Multi Topic Conference-IEEE INMIC 2001. Technology for the 21st Century*, pp. 223–228, Bahawalpur, Pakistan, November 2001.

[23] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, 2015.

[24] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pp. 71–80, Amsterdam, Netherlands, September 2012.

[25] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on twitter," :https://arxiv.org/abs/1706.01206, 2017.

[26] P. Mishra, M. Del Tredici, H. Yannakoudakis, and E. Shutova, "Abusive language detection with graph convolutional networks," 2019, https://arxiv.org/abs/1902.01748.

[27] R. Pelle, C. Alcântara, and V. P. Moreira, "A classifier ensemble for offensive text detection," in *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, pp. 237–243, Salvador, BA, Brazil, October 2018.

[28] J. M. Schneider, R. Roller, P. Bourgonje, S. Hegele, and G. Rehm, "Towards the automatic classification of offensive language and related phenomena in German tweets," in *14th Conference on Natural Language Processing, KONVENS 2018*, p. 95, Vienna, Austria, September 2018.

[29] A. Alakrot, L. Murray, and N. S. Nikolov, "Dataset construction for the detection of anti-social behaviour in online communication in Arabic," *Procedia Computer Science*, vol. 142, pp. 174–181, 2018.

[30] H. Rizwan, M. H. Shakeel, and A. Karim, "Hate-speech and offensive language detection in roman Urdu," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2512–2522, Hong Kong, China, November 2020.

[31] W. Abbas, *Toxic Comment Classification of Roman Urdu Text*, Department of Computer Science, COMSATS University, Islamabad, Pakistan, 2019.

[32] S. Kausar, B. Tahir, and M. A. Mehmood, "ProSOUL: a framework to identify propaganda from online Urdu content," *IEEE Access*, vol. 8, 2020.

[33] R. Khan, A. Ahmad, A. O. Alsayed, M. Binsawad, M. A. Islam, and M. Ullah, "QuPiD attack: machine learning-based privacy quantification mechanism for PIR protocols in health-related web search," *Scientific Programming*, vol. 2020, Article ID 8868686, 11 pages, 2020.