IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 December 2024, accepted 18 January 2025, date of publication 27 January 2025, date of current version 30 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3534662

RESEARCH ARTICLE

An Automatic Approach for the Identification of Offensive Language in Perso-Arabic Urdu Language: Dataset Creation and Evaluation

SALAH UD DIN[1,2], SHAH KHUSRO[1], (Member, IEEE), FARMAN ALI KHAN[2], MUNIR AHMAD[3,4], (Senior Member, IEEE), OUALID ALI[5], AND TAHER M. GHAZAL[6], (Senior Member, IEEE)

[1]Department of Computer Science, University of Peshawar, Peshawar 25120, Pakistan
[2]Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock 43600, Pakistan
[3]Department of Computer Sciences, National College of Business Administration and Economics, Lahore 54000, Pakistan
[4]University College, Korea University, Seoul 02841, Republic of Korea
[5]College of Arts and Science, Applied Science University, Manama 5055, Kingdom of Bahrain
[6]Department of Networks and Cybersecurity, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman 19111, Jordan

Corresponding authors: Salah Ud Din (salah_ud_din@cuiatk.edu.pk) and Farman Ali Khan (farman_marwat@ciit-attock.edu.pk)

ABSTRACT Offensive language is a type of unacceptable language that is impolite amongst individuals, specific community groups, and society as well. With the advent of various social media platforms, offensive language usage has been widely reported, thus developing a toxic online environment that has real-life endangers within society. Therefore, to foster a culture of respect and acceptance, a prompt response is needed to combat offensive content. On the other hand, the identification of offensive language has become a challenging task, specifically in low-resource languages such as Urdu. Urdu text poses challenges because of its unique features, complex script, and rich morphology. Applying methods directly that work in other languages is difficult. It also requires exploring new linguistic features and computational techniques on a relatively large dataset to ensure the results can be generalized effectively. Unfortunately, the Urdu language got very limited attention from the research community due to the scarcity of language resources and the non-availability of high-quality datasets and models. This study addresses those challenges, firstly by collecting and annotating a dataset of 12020 Urdu tweets using OLID taxonomy as a benchmark. Secondly, by extracting character-level and word-level features based on bag-of-words, n-grams and TFIDF representation. Finally, an extensive series of experiments were conducted on the extracted features using seven machine learning classifiers to identify the most effective features and classifiers. The experimental findings indicate that word unigrams, character trigrams, and word TFIDF are the most prominent ones. Similarly, among the classifiers, logistic regression and support vector machine attained the highest accuracy of 86% and F1-Score of 75%.

INDEX TERMS Offensive language identification, Urdu language dataset, OLID taxonomy, machine learning classifiers, cyberbullying, hate speech, profanity.

I. INTRODUCTION

Social media platforms, such as X, Facebook, Instagram, and YouTube, facilitate the widespread sharing of information, opinions, and experiences worldwide [1]. The global usage of social media platforms is expected to reach 429 million users by the end of 2024, and this figure is projected to reach 503 million by the year 2028.[1] The social media platform X[2] (formerly known as Twitter) was launched in 2006. The current active users of this platform are more than 400 million. It has a micro-blogging structure that enables users to

The associate editor coordinating the review of this manuscript and approving it for publication was Roozbeh Razavi-Far.

[1]https://www.statista.com/
[2]https://x.com/

participate in ongoing discussions and also informs users about new developments, if any, through issuing notifications. Over time, X has emerged as a significant instrument, playing a key role in highlighting and promoting domestic as well as international politics. This platform is now used by most of the world governments to advocate their policies and interact with citizens. Twitter appeals to a diverse range of users, hailing from different backgrounds, speaking various languages, and belonging to myriad religious affiliations. It gives users the privilege to articulate thoughts and views without any censorship or control [2]. Censoring content on social media restricts freedom of speech and hampers individuals' capacity to express themselves. However, users exploit this liberty over time. As a result of this, the offensive content on such platforms is increasing day by day.

A text containing ''any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct.'' is called offensive language [3]. Usually, such content includes comments about racism, sexism, gender, religion, hate speech, bullying, toxic, and other personal characteristics [4]. The effectees of such content go through emotional distress, heightened anxiety, and in some cases lead individuals to take their own lives. For example, Author [5] reported that individuals who experience cyberbullying are more likely to engage in suicidal behavior. Offensive language not only creates a toxic online environment but also has real-life consequences, as it fuels offline acts of violence, reinforces harmful stereotypes, and deepens societal divisions [6]. Therefore, it is crucial for individuals, social media platforms, and the society to actively combat offensive content and promote a culture of respect, acceptance, and inclusivity.

Offensive language identification is a challenging issue, and social media platforms are now actively searching for and eliminating such content by making use of reporting methods and algorithms based on machine learning. This has led to a significant amount of research to identify, prevent, and develop countermeasures with respect to offensive content. Social media platforms allow users to post their comments in international and local languages. The majority of the users prefer to post comments using their own native language. Even though social media is used in hundreds of languages and dialects around the world [7], but most of the research studies solely considered the English language. Due to the abundance of linguistic resources, the predominant focus of offensive language identification studies, such as [3], [8], [9], [10], lies in the English language. This is followed by the other renowned languages, e.g., German [10], [11], Italian [13], [14], Hindi [10], [15], and Arabic [1], [16]. Only a limited number of research studies have investigated low-resource languages such as Urdu [2], [17], Japanese, Indonesian, Spanish [18], Danish [19], Portuguese [20], Sinhala [21], Korean [22], Turkish [23], [24], Greek [25], Marathi [7] and Persian [26], [27].

Urdu, originating in the 12th century, belongs to the Indio-Aryan branch within the Indio-European family of languages. It originated from the fusion of Sanskrit and Persian in northern India. It was derived from the dialect spoken in the vicinity of Delhi and was significantly influenced by Arabic, Turkish, and English [26]. Urdu is written in Nastalīq script, a modified form of the Perso-Arabic script that follows a right-to-left direction. This script is derived from the Persian alphabet, which in turn is derived from Arabic [28]. Roman Urdu or Roman script is another script used to write Urdu using Latin alphabets. Urdu is the national language of Pakistan and the official language of both Pakistan and India. Nearly every nation on Earth has native speakers of this language. With an estimated 230 million speakers (including both native and non-native speakers), Urdu is the tenth most common language in the world, according to The World Factbook 2024.[3] Urdu is a low-resource language characterized by its unique syntax and morphological structure, which define it as distinct from Arabic and other resource-rich languages. It exhibits a flexible word order, allowing for several arrangements of words inside a single sentence. Words do not always have spaces between them, and letters might change shape based on their position within a word. Text can be written in a variety of styles, just as words with the same meaning can be written in various forms [29].

Offensive language identification in Urdu text poses challenges due to its unique features and it is not easy to adapt methods from other languages to Urdu. These features include a flexible word order, the absence of spaces between words, and the possibility of letter shape changes within a word. Additionally, a deep understanding of the language and significant pre-processing efforts are necessary to effectively utilize computational approaches in Urdu text. Regarding the offensive language identification in Urdu language, an initial study was conducted by Akhter et al. [17]. They developed a dataset of 2151 YouTube comments and evaluated various machine learning classifiers to identify the offensive language. In addition to this study, recently Hussain et al. [30] developed a balanced dataset of 7500 Facebook posts and classified it into offensive and non-offensive categories. In these studies, the researchers did not follow the dataset development guidelines commonly used for resource-rich languages. The dataset developed for the former study is the only available dataset to complement the existing research. Apart from these two preliminary studies [17], [30] research on Urdu offensive language suffers from the lack of classifiers and high-quality datasets. To generalize findings, it is needed to investigate new language features and computational approaches on a comparatively large dataset.

Therefore, the main objective of this research study is to complement the existing research on Urdu offensive language identification in general and develop a larger dataset based on well-defined guidelines, followed by the research community of resource-rich languages. The major contributions of this research work are: Firstly, to collect a huge collection of tweets for the dataset based on random sampling instead of biased sampling. In biased sampling [31], seed words or

[3]https://www.cia.gov/the-world-factbook/

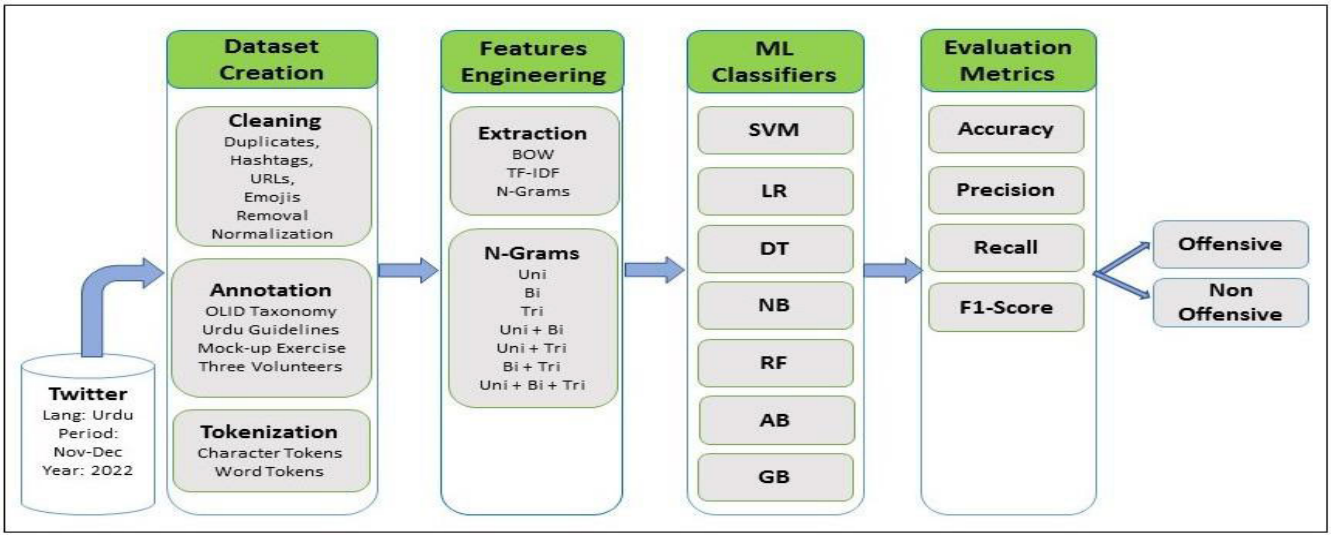

FIGURE 1. The methodology used in this study to identify the use of offensive language in urdu tweets.

lexicons are used during data collection which leads to the poor performance of classifiers on unseen data due to the presence of salient features. Secondly, to pre-process and annotate the dataset based on OLID taxonomy, well established definitions, and guidelines available in the literature. Thirdly, to explore the impact of various features on the performance of machine learning classifiers. Finally, to evaluate the classifiers performance on the newly developed dataset and analyze their results to find the best suited classifier for Urdu offensive language identification. The methodology followed for this study is presented in Figure 1.

The rest of a paper is structured as follows. Section II presents a literature review on various forms of offensive language identification in Urdu and Roman Urdu. Section III describes the definition and taxonomy of offensive language, along with the guidelines used during the data collection and annotation process for Urdu Offensive Language Dataset (UOLD). Lexical features, machine learning classifiers and evaluation metrics are illustrated in Section IV, while experimental setup, results, and findings of the study are discussed in Section V. Finally, Section VI concludes the paper with future work and limitations.

II. RELATED WORK

Offensive language is deeply connected to many social and linguistic issues, including hate, abuse, aggressiveness, bullying, and profanity. This section provides a brief overview of prior studies on offensive language and its associated issues in the Urdu language and Roman Urdu.

The first study to investigate the use of offensive language in the Urdu language was conducted by Akhter et al. [17]. They developed a balanced dataset from YouTube comments and evaluated various machine learning classifiers based on n-gram features. Logit Boost, a regression-based classifier that utilizes character trigram features, outperforms the other classifiers in terms of F1-score.

A similar study [32] was conducted to detect the use of abusive language by employing machine and deep learning classifiers on two datasets. The convolutional neural networks (CNN), as compared to other models achieved a high accuracy of 96.2% for the Urdu, and 91.4% for the Roman Urdu Script. Almost all of the applied models performed better over the Urdu dataset, and comparatively showed lower performance over the Roman Urdu dataset.

Recently, in another study [30] a dataset of 7.5k Urdu posts from Facebook was used to identify offensive and non-offensive posts. A seed words dictionary of Urdu offensive terms was used to collect these posts which are then annotated by human experts. An ensemble of logistic regression (LR), support vector machine (SVM), and stochastic gradient descent (SGD) classifiers utilizing word2vec features outperforms the individual classifiers. They also reported improvement in the results when the proposed ensemble classifier was employed on Urdu Offensive Dataset (UOD) [17].

Using Multilayer Perceptron (MLP), Ada-Boost (AB), Random Forest (RF), SVM, LR, and Long Short-Term Memory (LSTM), another study [2] investigated the task of finding threatening language in Urdu tweets. Researchers randomly chose a balanced dataset of 3.5k tweets, consisting of an equal number of threatening and non-threatening instances, from an imbalanced dataset of 10k tweets. MLP classifier with the word n-grams achieved the highest accuracy and F1-score of 72.5% and 72.7%, respectively, showing significant room for improvement.

A lexicon-based technique was proposed in one of the early studies [33] for detecting cyberbullying in Roman Urdu. A highly skewed dataset of 17K Roman Urdu tweets with only 379 instances of bullying was developed. A lexicon of tokens containing bullying tokens with negative scores and

non-bullying tokens with positive scores is used to annotate each tweet. A feature vector is created by assigning a label of ''bulled'' or ''not bulled'' to each tweet based on the lexically assigned score. The random forest classifier obtained superior performance as compared to Naive Bayes (NB), Decision Trees (DT), K-Nearest Neighbours (KNN), and SVM.

A lexicon of 621 words and a Roman Urdu dataset of 10k tweets with five classes were developed by the study's author [34] in order to categorize hate speech and offensive language. A CNN-gram model that excels on other deep learning models was introduced, along with a word embedding. it is pointed out that models struggle to distinguish between classifications of profanity, offensiveness, and abuse.

The findings of a study [35] that aimed to categorize Roman Urdu comments as offensive, hate speech, or neutral concluded that logistic regression with the count vector feature outperforms RF, SVM, DT, and CNN classifiers. These classifiers were assessed using a newly created dataset of 5K tweets with imbalanced classes. The dataset has been pre-processed to remove spelling variants.

A study conducted by researchers [36] found that SVM applied to n-grams for classifying offensive comments outperformed Logistic Regression and Naive Bayes. The author has developed a dataset consisting of 16k comments written in Roman Urdu. The author enhances the dataset using the Synthetic Minority Oversampling Technique (SMOTE) to address the issue of imbalanced classes.

One of the recent study [37] focused on the cyberbullying identification from Roman Urdu. A dataset of about 10k tweets containing five labels was evaluated using ML and Gated Recurrent Unit (GRU) classifiers. The GRU classifier utilizing lexical normalization outperformed the other classifiers in terms of accuracy and f1-score. Similarly another study [38] adopted a holistic approach for classifying online cyberbullying messages. They introduced a dataset that features user conversations in Urdu and Roman Urdu. The focus was on identifying repetition and intent to harm within the conversations. The approach was based on aggression detection using fine-tuned m-BERT and MuRIL classifiers, and they found MuRIL to be the most effective classifier. Furthermore, the approach adopted by [39] employed data augmentation through back-translation in order to double the samples of dataset. They reported an improvement of 8% in accuracy for detecting threatening language as compared to results reported by However, the LSTM classifier proposed by [2] them underperformed on the original dataset.

All the mentioned studies summarized in Table 1, show that most of the studies were focused on the Roman Urdu and various aspects of offensive language such as Hate Speech, toxicity, abuse, threat, and bullying. Moreover, only two preliminary studies investigated the problem of offensive language for Urdu.

III. URDU OFFENSIVE LANGUAGE DATASET (UOLD) CREATION

Offensive language identification is a text classification problem, and the effectiveness of research in this field relies on factors such as the quantity and quality of data, the selection of data samples, and the annotation schema employed for data labelling. This section presents the offensive language

TABLE 1. Summary of prior works on offensive language identification for urdu language including works on roman script of urdu language.

Study	Dataset	Dataset Available	Feature Engineering	Classifiers
YouTube [17]	Urdu (2.1K Comments) Roman Urdu (10K Comments)	Yes	Char n-grams, Word n-grams	NB, SVM, RF, JRip, Logit Boost, IBk
YouTube [32]	Urdu (2.1K Comments) Roman Urdu (10K Comments)	Yes	Bag-of-Word	CNN, LSTM, BLSTM, CLSTM, NB, SVM, IBK
Facebook [30]	Urdu (7,500 posts)	NO	Word N-grams, Bag Of words, TF-IDF, Word2vec	SVM, LR, SGD, RF, Ensemble of LR, SVM, SGD
Twitter [2]	Urdu (3.5K Tweets)	Yes	Char n-gram, Word n-gram, Text embeddings	LR, MLP, AB, RF, SVM, LSTM
Twitter [33]	Roman Urdu (17K tweets)	No	bag-of-words (BOW), trigram, Word2Vec	NB, SVM, DT, RF, KNN
Twitter [34]	Roman Urdu (10K tweets)	Yes	Unigram, Bigram, Trigram, Quad-gram.	CNN-gram
Twitter [35]	Roman Urdu (5K comments)	No	CV, TF, TF-IDF, Word2Vec	LR, RF, SVM, DT, CNN
YouTube [36]	Roman Urdu (16.3K comments)	No	Unigram, Bigram, Trigram, TF-IDF, SMOTE	SVM, LR, NB,

definition and its related taxonomy, along with the details of data collection, sampling, cleaning, and annotation process.

A. OFFENSIVE LANGUAGE: DEFINITION AND TAXONOMY

A well-defined annotation schema is a key aspect that defines offensive language and its subtypes to clearly reflect the problem [19]. Waseem et al. [40] pointed out that abusive language, offensive language, hate speech, and cyberbullying are partially overlapping phenomena. They also highlighted that there is a lack of consensus among the research community on definitions and relationships of these phenomena. This lack of consensus has resulted in contradictory guidelines for data annotation and incompatibility of developed datasets. They proposed a typology for defining abusive language and guidelines for data annotations, considering the definition contradictions. Later on, Zampieri et al. [3] effectively enhanced their proposed typology by capturing the target and type of offensive language. They concluded that any text containing "any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct," is offensive language. Based on this definition, three levels of hierarchical annotations, as shown in Figure 2 were suggested to classify the texts as non-offensive, profanity, cyberbullying, hate speech, or others.

At level A, text is examined for impolite, unpleasant, or non-acceptable language and annotated as offensive or non-offensive text. At level B, offensive text is examined for the presence of a target and annotated as either targeted or untargeted offensive text. Profanity refers to an offensive text without any target. At level C, a targeted offensive text is examined to investigate the type of target and annotated as individual, group, or other. Cyberbullying refers to an offensive text that targets a named individual or unnamed person. A group is formed by people who share the same religious beliefs, ethnicity, sexual orientation, political affiliation, gender, or common characteristic, and offensive text targeted at them is called hate speech. In addition to cyberbullying and hate speech, the "other" category includes offensive text that targets a situation, an issue, an event, or an organization.

A team of human annotators manually annotated tweets using this three-level schema and developed the Offensive Language Identification Dataset (OLID) of the English language for evaluation purposes. In their second attempt, they applied the same three-level OLID taxonomy in a semi-supervised way and developed SOLID, an expanded dataset of nine million tweets. Researchers working on offensive language identification in various languages have widely adopted the OLID taxonomy with its three-level schema. They have developed various datasets based on the OLID taxonomy, which serve as benchmarks for evaluating offensive language identification models in those languages. Table 2 provides a summary of the datasets developed based on the OLID taxonomy for different languages.

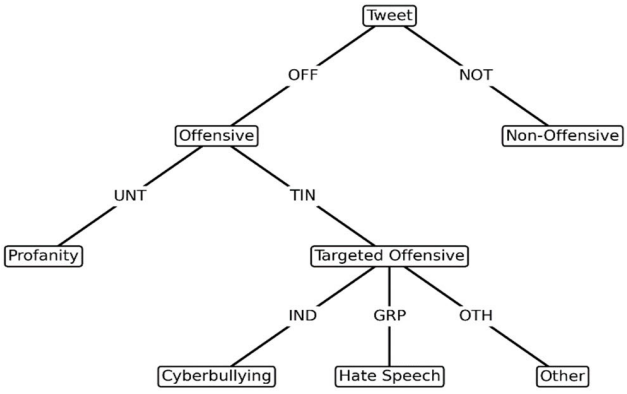

FIGURE 2. Three levels of hierarchical annotations based on OLID taxonomy.

B. DATA COLLECTION

In this study, OLID taxonomy is applied to develop the Urdu Offensive Language Dataset (UOLD). Initially, 2.5 million Urdu tweets were extracted from Twitter Using a custom Python script that utilized the Snscrape open-source library. Approximately 50K tweets per day from November to December 2022 are collected by setting the library language

TABLE 2. OLID Taxonomy-based datasets developed for various language.

Language	Platform	Size	Annotation
English (OLID) [3]	Twitter	14.1k	Humans
English (SOLID) [41]	Twitter	9.0M	Semi-supervised
Arabic [16]	Twitter	10K	Humans
Turkish [23]	Twitter	35.5K	Humans
Danish [19]	Facebook, Reddit, newspaper	3.6K	Humans
Greek (OGTD) [25]	Twitter	10.2K	Humans
Korean (KOLD) [22]	NAVER & YouTube	40K	Humans
Sinhala (*SOLD*) [21]	Twitter	10K	Semi-supervised
Marathi (MOLD) [7]	Twitter	3.6K	Humans
Persian (Pars-OFF) [[27]	Twitter	10.5K	Humans
Urdu (UOLD) [This Study]	Twitter	12K	Humans

filter to Urdu (lang: ur). Typically, the quantity of offensive tweets is significantly less than the quantity of non-offensive tweets, which makes the annotating process time-consuming. To address this issue, previous studies summarized in Table 2 employed a range of biased sampling techniques, including keyword searches, seed word lexicons, tracking reactions to common targets, and monitoring known users of offensive language throughout data collection. To ensure an unbiased dataset that accurately reflects the prevalence of offensive language, this study randomly collects tweets without employing any biased sampling technique.

C. DATA CLEANING

A day-wise uniform sample of 200K tweets is chosen from the initial collection for further pre-processing. Duplicate tweets, tweets with less than three words, and tweets containing unreadable characters are sampled out. Tweets that only contain URLs, hashtags, and emojis are also sampled because these types of tweets typically lack substantial text that can be analyzed for offensive language. Additionally, tweets written in languages other than Urdu, such as Pushto, Saraiki, Persian, Punjabi, Hindko, and Arabic, are excluded, as our study specifically focused on offensive language in Urdu tweets. Tweets published from the accounts of national and international organizations, public offices, political leaders, government officials, TV channels, and print and electronic media are discarded assuming that these accounts are unlikely to publish offensive content. After applying these pre-processing steps, a dataset of 12401 tweets is annotated by a team of three human annotators based on OLID taxonomy.

D. DATA ANNOTATION

The annotation process involved a team of three volunteers who were Urdu language experts and regular Twitter users. These annotators followed detailed guidelines written in both Urdu and English, which included examples of tweets to illustrate the criteria for annotation. The guidelines facilitated annotators to identify offensive language based on the OLID taxonomy. As a mock-up exercise, annotators were tasked to annotate a small number of selected tweets. We used the outcomes of this mock-up to discuss the mismatches, check the understanding of the guidelines, and highlight the common challenges associated with annotation. Inter-annotator agreement (IAA) is a crucial measure for the annotation process, as it provides a measure of the consistency and reliability of the annotations. In our study, we calculated IAA using Fleiss' coefficient, which considers the agreement among multiple annotators. A value of 61.5% for Fleiss' coefficient indicates a reasonable level of agreement among the three annotators, suggesting that the annotations are reliable and consistent.

A web interface-based application was developed and deployed on a local server to perform all the steps required for cleaning and annotation of tweets. The application interface shows tweets to annotator one by one, with the enhance navigational features. Annotator annotated each tweet by choosing a label from the dropdown list. The list contains available annotation labels. Application also included automated cleaning steps, such as removing URLs and normalizing usernames. Table 3 contains a few examples Urdu tweets and their labels by the team of annotators.

E. UOLD STATISTICS

The final UOLD dataset is an imbalance dataset, containing 8315 instances of non-offensive (69%) and 3705 instances of offensive (31%) tweets. Out of 8315 non-offensive tweets, 3705 tweets are randomly chosen to form a balanced UOLD dataset. For experiments, both imbalance and balanced dataset, is randomly divided into training and test sets by an 80:20 ratio such that the 69:31 ratio of non-offensive to offensive tweets is maintained. The statistics of both datasets are summarized in Table 4. By analyzing these datasets, we can gain a deeper understanding of the use of offensive language, its distribution, and its impact on online communication.

F. PRIVACY PROTECTION

Usernames mentioned in the tweet were replaced with @USER in the UOLD dataset to protect the privacy and anonymity of Twitter users.

IV. MATERIAL AND METHODS

Offensive language identification in Urdu text is a binary classification problem that can be solved using machine learning (ML) classifiers. ML classifiers like Logistic Regression (LR), Decision Trees (DT), Support Vector Machine (SVM), and Naïve Bayes (NB), while ensemble-based ML classifiers like Random Forest (RF), Gradient Boosting (GB) and Adaptive Boosting (AB) are chosen. These classifiers have shown promising results in text classification tasks [17], [30], [35], [36], making them suitable for tackling the complexities of offensive language identification in Urdu text. This section presents the essential description of selected classifiers, text representations utilized as features, and evaluation metrics such as accuracy, precision, recall, and the F1-score to evaluate classifiers' performance.

A. MACHINE LEARNING CLASSIFIERS

The LR classifier, based on the concept of probability, predicts the probability of a variable using the sigmoid function. This classifier uses features to construct a linear model using multinomial logistic regression. When using the LR classifier, the model calculates the probability of a text sample being offensive or non-offensive using the sigmoid function. Similarly, the SVM classifier finds support vectors (hyperplanes) and draws a linear or nonlinear decision boundary to classify the text input. SVM integrates a kernel function to analyze data patterns and transform data into higher-dimensional spaces to classify data points. The SVM classifier finds support vectors and draws a decision boundary to separate offensive and non-offensive text samples, utilizing a kernel function to capture complex patterns in the data. In contrast, the NB classifier uses conditional probability and

TABLE 3. A few example tweets from the UOLD dataset with their labels for three levels of OLID taxonomy.

No.	Tweet	A	B	C
1	دو تہائی کی اکثریت سے جیتنے کے خواب دیکھنے والا گھڑی چور نیازی آزاد کشمیر بلدیاتی انتخابات میں اسی تناسب سے مسترد ہوگیا	OFF	TIN	IND
2	جرنل حافظ عاصم منیر صاحب چھڑی دھو لینا یہ کئی ماوں کے لخت جگروں کے خون سے رنگی ہوئی ہے	OFF	TIN	IND
3	راستے میں ہی شلوار اتر گئ اب تو الٹی بھی پہن کر نہیں آ سکتا تھا۔ @USER	OFF	TIN	IND
4	باجوہ کی کمپنی میں ایک سے ایک بے غیرت بھرتی تھا ملک کی تباہی کے یہ سب زمہ دار @USER ہیں	OFF	UNT	_
5	یہ تنخواہ دار کٹھ پتلیاں اپنی لالچ کی وجہ سے بے بس ہیں، یہ بھی جانتے ہیں کہ ان کے آقا @USER کمینے، کرپٹ اور بے ایمان ہیں لیکن ان کی آنکھوں پر لالچی کے پردے ہیں، یہ وہ لوگ ہیں جو اپنے آقاؤں کی باقی ماندہ ہڈیوں پر پلے ہیں، یہ ہمیشہ غلام ہی رہیں گے۔	OFF	UNT	_
6	پی ایس ایف کا مقصد تعلیمی اداروں میں چرس ،سیگریٹ پینا لڑکیوں کو تنگ کرنا کینٹینوں میں مفت روٹیاں کھانا اور دکانوں سے چاکلیٹ چوری کرنا ان لوگوں کو طلباء کہنا طلبہ کی توہین ہے، مفت خوروں اور چاکلیٹ چوروں کا تعلیمی اداروں میں کوئی جگہ نہیں	OFF	TIN	OTH
7	ماں تو وہ اس کے بعد بھی پیش کریں گے کیونکہ عادت سے مجبور ہیں@USER	OFF	UNT	_
8	پی ٹی آئی واحد رنڈیوں کے بچوں کی جماعت ہے جس میں سارے دنیا کے رنڈیوں کے بچے شامل ہیں	OFF	TIN	OTH
9	قاتلو، ملک دشمنو، غدارو امید کی ایک کرن جاگی تھی کے غداری، قتل و غارت، اور فاشزم @USER کا یہ سلسلہ رکے گا، ہمیں نہ سہی ہماری نسلوں کو اللہ تم لوگوں کی کتوں سے بد تر موت ضرور دکھائے گا	OFF	TIN	OTH
10	انشاءاللہ ٹی ٹی پی کو چھپنے کی جگہ نہیں ملی گی-@USER یہ دجال کے بچے ہیں-	OFF	TIN	OTH
11	تاریخ اٹھا کر دیکھ لیں جس جس نے بھی پاکستان کو نقصان پہنچانے کی کوشش کی ان سب @USER کا ننانوے فیصد تعلق شیعوں یا قادیانیوں سے ضرور نکلتا ہے، وہ کوئی محکمہ ہو یا کوئی کلیدی عہدہ یہ دو گروہ بہرحال پاکستانیوں کو ہر ممکن عذاب میں ضرور ڈالتے ہیں-	OFF	TIN	GRP
12	آپ 10 کافروں کو مسلمان بنا سکتے ہیں مگر آپ 1 یوتھیئے کو انسان نہیں بنا سکتے	OFF	TIN	GRP
13	اوئے تم کو ارشد شریف نظر نہیں آتا.. آج اس کے قتل کی ایف آئی آر درج ہوئی اسکے @USER لواحقین کے بغیر یہاں کیوں نہیں بولتے...؟	NOT	_	_
14	رفیع بھائی آپ قادیانیوں کے بارے میں بات کرتے اتنے خوفزدہ ہیں کہ ایک فقرے میں تین @USER دفعہ آپ کو یہ کہنا پڑا کہ مجھے کوئی ہمدردی نہیں۔تو آپ سوچیں کہ قادیانی پاکستان میں کن حالات میں زندگی گزار رہے ہیں--اور آج تک کسی قادیانی پر کوئی الزام ثابت نہیں ہو سکا..	NOT	_	_
15	آج تاریخی بابری مسجد کے انہدام کی 30 ویں برسی ہے۔ ہندو انتہا پسندوں کی طرف سے نظر ثانی کا عمل اب ہندوستانی مسلمانوں اور دیگر اقلیتوں کے لیے ایک زندہ ڈراؤنا خواب بن گیا ہے۔ دنیا کو بھارت میں بڑھتی ہوئی مذہبی منافرت کا نوٹس لینا چاہیے۔وزیراعظم شہباز شریف	NOT	_	_
16	باجی آپ مشہوری چاہیتی ہیں@USER	NOT	_	_
17	اگر مرد پابندی لگاتا ہے تو اس کا مطلب یہ نہیں کہ اسے خاتون پر اعتماد نہیں بلکہ یہ کہ اسے دوسرے مردوں پر اعتماد نہیں---اگر خاتون سوال کرتی ہے تو اس لیے نہیں کہ وہ مرد پر شک کرتی ہے بلکہ اس لیے کہ اسے دوسری عورتوں پر اعتماد نہیں۔	NOT	_	_
18	آج آپ نے جرمنی کے جیتنے کی دعا کے ساتھ جاپان کے برابر رہنے یا ہارنے کی بھی دعا @USER کرنی تھی	NOT	_	_
19	مجھے سمجھ نہیں آتی کہ یہ جیالے اور پٹواری ہمیشہ کھانے پر کیوں لڑتے ہیں@USER	NOT	_	_
20	یہ توفیق بعض لوگوں کو اللہ تعالی دیتا ہر کسی کو نہیں@USER	NOT	_	_

TABLE 4. Summary of tweets distribution in imbalance and balanced UOLD dataset.

Dataset	Distribution	Non-Off (69%)	Off (31%)	Total
Imbalance UOLD	Train	6652	2964	9616 (80%)
	Test	1663	0741	2404 (20%)
	Total	8315	3705	12020
Balanced UOLD	Train	2964	2964	5928 (80%)
	Test	0741	0741	1482 (20%)
	Total	3705	3705	7410

TABLE 5. Results of classifiers using character and word BOW features.

Feature	Evaluation	LR	DT	MNB	SVM	AB	GB	RF	Best Classifiers
Char BOW	Accuracy	68	64	68	67	**72**	**72**	**73**	Accuracy RF, GB, AB F1-Score SVM, LR, MNB
	Precision	48	41	48	48	60	65	72	
	Recall	65	41	52	68	32	23	20	
	F1-Score	**55**	41	**50**	**56**	41	34	32	
Word BOW	Accuracy	**85**	78	**84**	**84**	82	80	83	Accuracy LR, SVM, MNB F1-Score LR, SVM, MNB
	Precision	77	64	80	76	82	89	91	
	Recall	72	64	65	68	55	41	50	
	F1-Score	**74**	64	**72**	**72**	65	56	64	

TABLE 6. Results of classifiers using character and word TFIDF features.

Feature	Evaluation	LR	DT	MNB	SVM	AB	GB	RF	Best Classifiers
Char TF-IDF	Accuracy	67	65	69	69	**73**	**74**	**73**	Accuracy GB, RF, AB F1-Score SVM, LR, MNB
	Precision	48	43	42	50	60	68	70	
	Recall	70	40	60	73	37	29	19	
	F1-Score	**57**	42	**50**	**59**	46	40	30	
Word TF-IDF	Accuracy	**84**	74	80	**86**	**81**	80	81	Accuracy SVM, LR, 81 F1-Score LR, SVM, AB
	Precision	75	58	86	84	80	89	92	
	Recall	73	60	44	65	52	41	44	
	F1-Score	**74**	59	58	**73**	**63**	56	59	

the Bayes theorem. NB makes the learning assumption that every feature within a given dataset is independent of one another. Consequently, each feature makes an equivalent contribution to the calculation of the predicted label. As a result of this supposition, NB is both efficient and straightforward to implement for large datasets. Despite this, the DT classifier uses a tree-like structure to classify data into different classes. DT uses training data to capture descriptive decision-making knowledge which defines a set of rules to classify the test data. A single decision tree may not capture complex relationships, overfit the training data, and favor dominant classes in case of an imbalanced dataset. Combining multiple decision trees (e.g., Random Forests or Gradient Boosting) can address these weaknesses.

B. ENSEMBLE-BASED ML CLASSIFIERS

Ensemble-based classifiers combine several weak classifiers to make the most flexible classifier. The RF ensemble classifier can leverage the collective knowledge of multiple decision trees to improve classification accuracy and prevent overfitting. Each decision tree is classified separately based on the rules derived from the selected group of features. The final classification is obtained through aggregation, often using a majority vote. The RF's novel structure allows each tree to offer different classification perspectives, preventing overfitting. On the other hand, an AdaBoost classifier is a meta-estimator that creates an ensemble of weak learners. It adjusts the weights of multiple weak classifiers into a strong classifier to improve the classification performance. Similarly, the GB classifier combines many weak decision trees in series to come up with one strong learner. This classifier uses the adaptive boosting method combined with weighted minimization to minimize the loss.

C. FEATURES ENGINEERING

The performance of classifiers largely depends upon the text representation used. Text representation, an essential step for any text classification task, is the process of transforming text into features. In this study, bag-of-word (BOW), term frequency-inverse document frequency (TF-IDF), and n-gram representation of both characters and words are utilized for text representation. The selection of n-grams, BOW, and TF-IDF as feature representations is based on their ability to capture different aspects of the text. BOW represents the frequency of words/characters in a document without considering the order of words, providing a simple and efficient representation. TF-IDF considers the importance of words/characters in a document based on their frequency in the document and across the corpus. Char n-grams are sequences of n-characters that help to capture character-level

TABLE 7. Results of classifiers using character n-grams features.

Chars	Evaluation	LR	MNB	DT	SVM	AB	GB	RF	Best Classifiers
Uni	Accuracy	72	67	72	73	**73**	**74**	**74**	Accuracy GB, RF, AB F1-Score GB, AB, SVM
	Precision	60	46	64	62	59	63	75	
	Recall	23	47	17	32	35	36	21	
	F1-Score	34	47	27	**42**	**44**	**46**	33	
Bi	Accuracy	**82**	78	74	**81**	78	**83**	79	Accuracy GB, LR, SVM F1-Score LR, GB, SVM
	Precision	76	69	66	75	73	81	94	
	Recall	62	51	32	60	61	58	33	
	F1-Score	**69**	59	43	**67**	67	**68**	49	
Tri	Accuracy	**86**	82	75	85	**85**	**86**	85	Accuracy LR, GB, AB F1-Score LR, GB, AB
	Precision	83	68	70	81	80	85	91	
	Recall	69	78	31	66	68	65	57	
	F1-Score	**75**	72	43	73	**73**	**74**	70	
Uni + Bi	Accuracy	**83**	76	74	81	**83**	**83**	78	Accuracy LR, AB, GB F1-Score LR, AB, GB
	Precision	76	62	66	67	76	83	91	
	Recall	63	57	29	74	64	57	30	
	F1-Score	**69**	59	40	71	**69**	**68**	45	
Uni + Tri	Accuracy	**86**	81	74	**85**	84	**85**	83	Accuracy LR, SVM, GB F1-Score SVM, LR, MNB
	Precision	82	65	70	76	83	86	92	
	Recall	69	80	28	75	63	60	51	
	F1-Score	**74**	**72**	40	**76**	71	71	65	
Bi + Tri	Accuracy	**86**	81	72	**85**	85	**86**	83	Accuracy LR, AB, SVM F1-Score SVM, AB, LR
	Precision	84	66	71	75	80	89	91	
	Recall	67	80	15	77	69	61	48	
	F1-Score	**74**	72	25	**76**	**74**	72	63	
Uni + Bi + Tri	Accuracy	**85**	80	73	**85**	84	**86**	82	Accuracy GB, SVM, LR F1-Score SVM, GB, LR
	Precision	83	79	66	75	81	86	89	
	Recall	67	49	28	77	62	66	47	
	F1-Score	**74**	60	40	**76**	71	**75**	61	

patterns and dependencies. Word n-grams are sequences of n-words that focus on word-level patterns and relationships. Char-based features can capture character-level patterns and dependencies, while word-based features focus on word-level patterns and relationships. By using text representations, we aim to capture both local and global textual features, enabling a comprehensive analysis of offensive language in Urdu text.

D. EVALUATION METRICS

The ability of classifiers to identify offensive language is assessed by the following evaluation measures. Accuracy measures the overall correctness of the classifier's predictions, providing an overall assessment of its performance. Precision focuses on the classifier's ability to correctly identify offensive language, measuring the proportion of true positives out of all positive predictions. Recall, also known as sensitivity, measures the classifier's ability to capture offensive language comprehensively by calculating the proportion of true positives out of all actual offensive instances. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of both metrics and considering both metrics to evaluate the classifier's performance.

V. EXPERIMENTAL SETUP AND RESULTS

Several experiments were designed to gain a deeper understanding of various features and classifiers to identify the use of offensive language in Urdu. Logistic Regression (LR), Decision Tree (DT), Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), AdaBoost (AB), Gradient

TABLE 8. Results of classifiers using word n-grams features.

Word	Evaluation	LR	MNB	DT	SVM	AB	GB	RF	Best Classifiers
Uni	Accuracy	**84**	**84**	71	**85**	84	84	84	Accuracy SVM, MNB, LR F1-Score MNB, SVM, LR
	Precision	75	76	68	81	80	80	89	
	Recall	70	72	11	67	65	63	56	
	F1-Score	**73**	**74**	19	**73**	72	71	68	
Bi	Accuracy	**78**	**80**	70	77	**77**	77	76	Accuracy MNB, LR, Ada F1-Score
	Precision	71	72	74	71	72	71	0.8	
	Recall	48	55	06	45	44	43	32	
	F1-Score	**57**	**62**	11	55	**55**	53	46	MNB, LR, Ada
Tri	Accuracy	**74**	**74**	69	**73**	73	73	72	Accuracy LR, MNB, SVM F1-Score RF, LR, MNB
	Precision	66	70	67	68	67	68	58	
	Recall	32	28	01	24	25	23	35	
	F1-Score	**43**	**40**	01	36	36	35	**44**	
Uni + Bi	Accuracy	**84**	**84**	72	84	**84**	84	82	Accuracy MNB, LR, Ada F1-Score MNB, LR, Ada
	Precision	79	80	68	77	80	82	91	
	Recall	67	65	18	67	67	61	47	
	F1-Score	**72**	**72**	29	72	**72**	70	62	
Uni + Tri	Accuracy	**85**	**84**	71	**85**	84	84	83	Accuracy SVM, LR, MNB F1-Score SVM, LR, MNB
	Precision	81	76	73	82	80	80	91	
	Recall	66	67	07	65	65	63	48	
	F1-Score	**73**	**72**	13	**73**	71	71	63	
Bi + Tri	Accuracy	**78**	**79**	71	**78**	77	77	76	Accuracy MNB, LR, SVM F1-Score MNB, LR, SVM
	Precision	76	74	81	72	72	78	86	
	Recall	44	49	06	44	43	36	27	
	F1-Score	**56**	**59**	11	**55**	54	50	42	
Uni + Bi + Tri	Accuracy	**85**	83	72	**84**	84	**85**	82	Accuracy LR, GB, SVM F1-Score LR, SVM, GB
	Precision	79	83	63	78	79	82	91	
	Recall	67	57	26	68	66	64	45	
	F1-Score	**73**	68	37	**73**	72	**72**	61	

Boosting (GB) and Random Forest (RF) classifiers were used in these experiments. The results and analysis of these experiments are presented. Experiments are performed using Scikit-learn and ThunderSVM Python libraries. The 'Best Classifiers' column in the result tables shows the top three classifiers based on accuracy and f1-score. These experiments were executed on two desktop computers, an HP Corei7, having 16GB memory, and an ASUS Z97-K, having NVIDIA 1060 GPU, 32GB CPU memory, and 6GB GPU memory. The dataset, along with the source code, is available online.[4]

A. RESULTS USING BOW FEATURES

BOW is a very simple and fundamental feature that represents the frequency of characters or words in a tweet without accounting for the word order. In this experiment, character unigrams and word unigrams were extracted from the UOLD dataset, and their counts were passed on to the classifiers as BOW features. The classifiers were trained with default parameters, and their results are shown in Table 5. For character-based features, the performance of RF, GB, and AB classifiers was noteworthy, with an accuracy of 73%,72%, and 72%, respectively. Similarly, with respect to the F1-Score, SVM, LR, and MNB attained the highest performance. On the other hand, for the word-based features, the performance of LR, SVM, and MNB classifiers was significant with respect to both accuracy and F1-score. The accuracy of the mentioned classifiers was at 85%, 84%, and 84%, respectively. Therefore, based on the results, the LR classifier with the word feature was the most effective, and the DT classifier was the least effective classifier. Additionally, word BOW features are more significant than character BOW features.

[4]https://github.com/salah-ud-din/Urdu-Offensive-Language-Identification

TABLE 9. Results of classifiers using character n-grams features from balanced UOLD dataset.

Feature	Evaluation	LR	MNB	DT	SVM	AB	GB	RF	Best Classifiers
Uni	Accuracy	67	64	63	**68**	**68**	68	**69**	Accuracy RF, SVM, AB F1-Score SVM, RF, AB
	Precision	67	61	61	65	69	68	69	
	Recall	65	77	74	77	67	68	71	
	F1-Score	66	68	67	**71**	**68**	68	**70**	
Bi	Accuracy	**79**	74	63	78	**78**	**80**	77	Accuracy GB, LR, AB F1-Score GB, LR, AB
	Precision	80	71	63	80	79	81	75	
	Recall	77	80	62	73	77	79	80	
	F1-Score	**78**	75	63	77	**78**	**80**	78	
Tri	Accuracy	**84**	81	82	**83**	82	**84**	74	Accuracy GB, LR, SVM F1-Score GB, LR, SVM
	Precision	88	80	84	88	85	87	78	
	Recall	78	83	80	77	78	79	68	
	F1-Score	**83**	82	82	**82**	81	**83**	73	
Uni + Bi	Accuracy	**79**	71	76	**79**	78	**80**	66	Accuracy GB, LR, SVM F1-Score GB, LR, SVM
	Precision	80	68	74	80	77	82	65	
	Recall	77	80	79	77	78	77	66	
	F1-Score	**79**	73	76	**78**	78	**79**	66	
Uni + Tri	Accuracy	**83**	81	80	**83**	82	**83**	72	Accuracy GB, LR, SVM F1-Score GB, LR, SVM
	Precision	86	80	80	86	84	85	73	
	Recall	78	84	80	78	79	80	70	
	F1-Score	**82**	82	80	**82**	81	**83**	72	
Bi + Tri	Accuracy	**83**	80	81	83	83	**84**	71	Accuracy GB, LR, SVM F1-Score GB, LR, SVM
	Precision	86	79	82	85	85	87	73	
	Recall	79	82	81	81	81	80	68	
	F1-Score	**83**	80	81	83	83	**83**	70	
Uni + Bi + Tri	Accuracy	**83**	79	80	**83**	81	**83**	70	Accuracy GB, LR, SVM F1-Score GB, LR, SVM
	Precision	86	77	80	85	83	84	70	
	Recall	79	83	80	79	79	80	72	
	F1-Score	**82**	80	80	**82**	81	**82**	71	

B. RESULTS USING TFIDF FEATURES

TF-IDF is an effective feature that captures the frequency of characters or words in a tweet, as well as across the entire dataset. In this experiment, TF-IDF was computed for both character and word unigrams from the UOLD dataset and used as input for classifiers with default parameters. The classification results are presented in Table 6. For character-based features, the GB, RF, and AB classifiers were the top performers with accuracies of 74%,73%, and 73%, while SVM, LR, and MNB classifiers were the top scorers in terms of F1 scores. In the case of word features, the highest accuracies of 86%,84%, and 81%, and F1-scores of 73%, 74%, and 63% were achieved by SVM, LR, and AB classifiers, respectively. From these results, we can conclude that the SVM is the most effective classifier in terms of accuracy, and LR is most effective in terms of the F1-score for the case of TF-IDF features. Furthermore, the word TF-IDF proves to be more significant than the character TF-IDF.

C. RESULTS USING N-GRAM FEATURES

n-grams are the sequences of n-characters or n-words that capture the patterns in text. Character n-grams capture character-level patterns, while word n-grams focus on word-level relationships. For this experiment, character-level and word-level unigrams, bigrams, and trigrams were computed from the UOLD dataset. These unigrams, bigrams, trigrams, and their combinations were then provided as input to classifiers. A randomized search with 5-fold cross-validation was applied during training to optimize parameters and validate the results. The performance of classifiers using character

TABLE 10. Results of classifiers using word n-grams features from balanced UOLD dataset.

Word	Evaluation	LR	MNB	DT	SVM	AB	GB	RF	Best Classifiers
Uni	Accuracy	81	**82**	70	81	81	**81**	**82**	Accuracy MNB, RF, GB F1-Score MNB, SVM, RF
	Precision	85	81	71	83	84	86	85	
	Recall	76	83	65	78	77	75	77	
	F1-Score	80	**82**	68	**81**	80	80	**81**	
Bi	Accuracy	**72**	**74**	64	**71**	67	68	71	Accuracy MNB, LR, SVM F1-Score MNB, LR, SVM
	Precision	73	77	66	78	72	69	80	
	Recall	69	70	55	60	56	65	56	
	F1-Score	**71**	**73**	60	**68**	63	67	66	
Tri	Accuracy	**65**	**65**	62	**65**	60	61	64	Accuracy MNB, LR, SVM F1-Score LR, SVM, RF
	Precision	68	74	67	69	71	72	69	
	Recall	55	45	47	53	35	36	50	
	F1-Score	**61**	56	56	**60**	47	48	**58**	
Uni + Bi	Accuracy	**81**	**81**	69	81	79	**82**	81	Accuracy GB, LR, MNB F1-Score LR, MNB, GB
	Precision	83	81	70	83	88	87	88	
	Recall	78	79	68	79	68	75	72	
	F1-Score	**81**	**80**	69	80	76	**80**	79	
Uni + Tri	Accuracy	**82**	81	71	**81**	**82**	81	75	Accuracy LR, SVM, AB F1-Score LR, SVM, AB
	Precision	85	83	74	84	85	86	82	
	Recall	78	79	66	78	78	74	65	
	F1-Score	**81**	81	70	**81**	**81**	80	72	
Bi + Tri	Accuracy	**72**	**74**	66	**71**	68	68	71	Accuracy MNB, LR, SVM F1-Score MNB, LR, SVM
	Precision	76	77	69	75	71	76	80	
	Recall	65	69	58	62	61	51	56	
	F1-Score	**70**	**73**	63	**68**	66	61	66	
Uni + Bi + Tri	Accuracy	**81**	80	69	**81**	**81**	81	80	Accuracy LR, SVM, AB F1-Score LR, SVM, AB
	Precision	84	82	71	84	84	86	88	
	Recall	77	77	65	77	76	74	69	
	F1-Score	**80**	80	68	**80**	**80**	79	77	

n-grams is presented in Table 7. The LR, AB, and SVM classifiers were the top performers in terms of accuracy, while SVM, LR, and GB excelled in terms of F1 scores. The LR classifier, using character trigrams, achieved the highest accuracy at 86% and an F1 score of 75%. Similarly, the ensemble classifier AB reached the same performance using a combination of unigrams, bigrams, and trigrams. The SVM classifier followed, with an accuracy of 85% and F1-score of 75%. For word n-grams, both LR and SVM classifiers performed equally well, as depicted in Table 8. These classifiers excelled in the combination of unigrams and trigrams, achieving an accuracy of 85% and an F1 score of 73%.

The results from Table 7 and Table 8 also provide a comparison between character and word n-grams. For character n-grams, classifiers performance improves as the n-grams progress from unigrams to trigrams. In contrast, for word n-grams, classifier performance decreases as the n-grams move from unigrams to trigrams. Additionally, the performance of classifiers on combined n-grams generally matches the highest performance achieved with individual n-grams within the combination. Based on these findings, we can infer that Logistic Regression is the most effective classifier for n-grams, with character n-grams proving to be more significant than word n-grams.

D. COMPARISON WITH PRELIMINARY STUDY RESULTS

In this experiment, n-gram features were extracted from the balanced UOLD dataset to compare our results with those of the preliminary study [17]. The findings of the preliminary study indicate that the character trigrams are the most significant feature and that the LR classifier is the best performer, with an F1 score of 95.9%. Tables 9 and 10 present the results of our experiment, showing that the LR classifier using character trigrams achieved the highest accuracy and F1-score of 84% and 83%, respectively. Similarly, the MNB classifier using word unigrams attained an accuracy and F1-score of 82%. This concludes that the findings of our experiment for the LR classifier and character trigrams agree with the preliminary study findings. However, the F1-score of the LR classifier in our experiment achieved a lower value. In this

TABLE 11. Performance and runtime comparison of top performing classifiers.

Classifier	Feature	Accuracy	Precision	Recall	F1-Score	Training (in sec)	Test (in sec)
RF	Char Bow	73	71	20	31	9.43	0.25
LR	Word BOW	85	77	72	74	26.18	0.39
GB	Char TF-IDF	74	68	29	40	13.62	0.02
SVM	Word TF-IDF	85	80	68	73	14754.31	884.51
LR	Char tri	86	77	77	77	3.05	0.12
SVM	Word Uni	85	77	74	75	1394.80	95.11

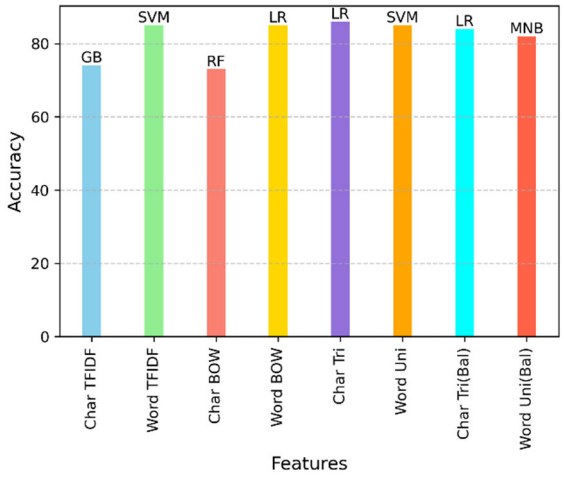

FIGURE 3. Accuracies of chosen classifiers on various features. The results on features extracted from a balanced dataset are indicated with 'Bal.'

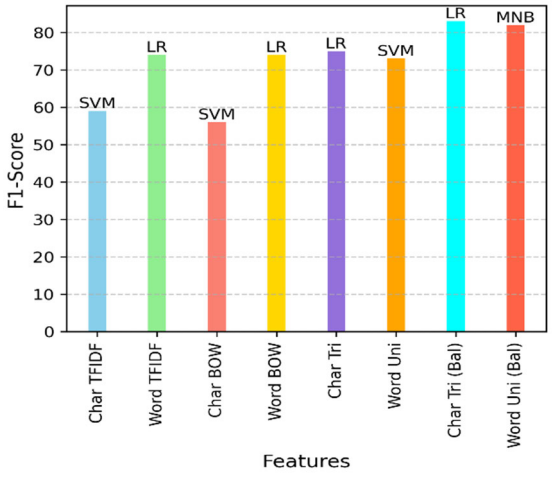

FIGURE 4. F1-Scores of chosen classifiers on various features.

regard, the study [32] reported that classifiers tend to perform lower on larger datasets than on smaller datasets. Therefore, the reason behind this lower score is the linguistic variations and expanded vocabulary of our dataset.

E. RESULTS DISCUSSION

The results of all the mentioned experiments indicate the performance of classifiers on various individual features such as BOW, TF-IDF, and n-grams. The top-performing classifiers of these experiments are chosen to analyze the performance of the classifiers across these features. Figure 3 and Figure 4 compare the performance of these chosen classifiers, evaluated on accuracy and F1-score metrics. The LR and SVM classifiers, fed respectively with character trigrams and word TFIDF features from the imbalance dataset, show the highest accuracy of 86% and F1-Score of 75%. The RF and SVM classifiers fed respectively with character BOW and TFIDF features attained the lowest accuracy of 73% and F1-Score of 59%. On the other hand, the LR classifier fed with character trigram features extracted from a balanced dataset attained the highest F1-Score of 83%. Results also indicate that the LR classifier shows consistent performance across different features, and SVM shows variable performance, with high accuracy in word TFIDF but lower F1-scores in character TFIDF and BOW features. The LR classifier, among all the features, appears to provide the best balance of accuracy and F1 score.

The top performing classifiers of all experiments, their training and testing time (in second) are shown in the Table 11. The time complexity of LR is O(Nd), RF is O(TN logN), GB is O(TN logN), and kernelized SVM is O(N3). While the memory complexity of LR is O(d), RF is O (Tdn), GB is O(Tdn), and kernelized SVM is O(N2). Here, N is the number of tweets, d is the number of features (134 for character features, 9,744 for word unigrams, 15,8741 for character trigrams, and 25,145 for word features), T is the number of trees, and n is the number of nodes. The LR classifier exhibited the lowest runtime and memory cost, RF and GB incurred intermediate costs, and SVM had the highest computational and memory cost, as summarized in Table 11.

VI. CONCLUSION AND FUTURE WORK

In this study, an Urdu offensive language dataset (UOLD) is developed to identify the use of offensive language in Urdu tweets. Urdu has unique features compared to high-resource languages and is much less focused on by the research community due to the absence of high-quality datasets. We addressed this issue by collecting tweets without any biased sampling and annotating them by following OLID taxonomy. UOLD is an imbalanced dataset containing

non-offensive and offensive tweets in a ratio of 69:31. Both word level and character level features, such as bag-of-word, term frequency, inverse document frequency, and n-grams, are extracted as features. These individual and combinations of these individual features are passed to seven machine learning classifiers to identify the presence of offensive language. The effectiveness of individual and combined features on the performance of classifiers is also investigated and compared. The logistic regression and support vector machine classifiers showed superior performance, scoring an accuracy of 86% and F1-Score of 75%, while the random forest classifier attained the lowest scores. Character trigrams, word unigrams, and word TFIDF are the most prominent features for the task of Urdu tweet classification. We believe that these contributions yield to enhance the ability of social media platforms with respect to effectively combating offensive content in Urdu.

In the future, we aim to extend the UOLD dataset by increasing the number of tweets to effectively incorporate the higher levels of OLID taxonomy. We also aim to apply neural network and deep learning classifiers like LSTM, RNN and Transformers on the UOLD dataset to perform the Level B and Level C classification of tweets.

REFERENCES

[1] S. Almutiry and M. A. Fattah, "Arabic CyberBullying detection using Arabic sentiment analysis," *Egyptian J. Lang. Eng.*, vol. 8, no. 1, pp. 39–50, Apr. 2021, doi: 10.21608/ejle.2021.50240.1017.

[2] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, and A. Gelbukh, "Threatening language detection and target identification in Urdu tweets," *IEEE Access*, vol. 9, pp. 128302–128313, 2021, doi: 10.1109/ACCESS.2021.3112500.

[3] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," 2019, *arXiv:1902.09666*.

[4] T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, and M. Granitzer, "I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language," in *Proc. 12th Lang. Resour. Eval. Conf.*, May 2020, pp. 6193–6202.

[5] U. Azam, H. Rizwan, and A. Karim, "Exploring data augmentation strategies for hate speech detection in Roman Urdu," in *Proc. 13th Lang. Resour. Eval. Conf.*, Jun. 2022, pp. 4523–4531.

[6] *Hate Speech on Social Media: Global Comparisons | Council on Foreign Relations*. Accessed: Jul. 20, 2024. [Online]. Available: https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

[7] M. Zampieri, T. Ranasinghe, M. Chaudhari, S. Gaikwad, P. Krishna, M. Nene, and S. Paygude, "Predicting the type and target of offensive social media posts in Marathi," *Social Netw. Anal. Mining*, vol. 12, no. 1, p. 77, Dec. 2022, doi: 10.1007/s13278-022-00906-8.

[8] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "SemEval-2019 task 6: Identifying and categorizing offensive language in social media (Offenseval)," 2019, *arXiv:1903.08983*.

[9] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin, "SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval)," 2020, *arXiv:2006.07235*.

[10] T. Mandl. *Hate Speech and Offensive Content Identification in Indo-European Languages*. Accessed: Dec. 16, 2023. [Online]. Available: https://www.academia.edu/download/93493898/417955443.pdf

[11] J. Risch, A. Stoll, L. Wilms, and M. Wiegand, "Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments," in *Proc. GermEval Shared Task Identificat. Toxic, Engaging, Fact-Claiming Comments*, 2021, pp. 1–12. Accessed: Dec. 16, 2023. [Online]. Available: https://aclanthology.org/2021.germeval-1.1/

[12] M. Wiegand, M. Siegel, and J. Ruppenhofer, "Overview of the GermEval 2018 shared task on the identification of offensive language," in *Proc. 14th Conf. Natural Lang. Process.*, 2018, pp. 1–10.

[13] S. Pelosi, P. Vitale, A. Maisto, and S. Vietri, "Mining offensive language on social media," in *Proc. 4th Italian Conf. Comput. Linguistics (CLIC-IT)*, R. Basili, M. Nissim, and G. Satta, Eds., 2017, pp. 252–256. Accessed: Dec. 16, 2023. [Online]. Available: https://www.iris.unisa.it/handle/11386/4717098

[14] C. Bosco, D. Felice, F. Poletto, M. Sanguinetti, and T. Maurizio, "Overview of the evalita 2018 hate speech detection task," in *Ceur Workshop Proc.*, 2018, pp. 1–9. Accessed: Dec. 16, 2023. [Online]. Available: https://iris.unito.it/bitstream/2318/1686264/1/paper010.pdf

[15] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking aggression identification in social media," in *Proc. 1st Workshop Trolling, Aggression Cyberbullying*, 2018, pp. 1–11. Accessed: Dec. 16, 2023. [Online]. Available: https://aclanthology.org/W18-4401/

[16] H. Mubarak, A. Rashed, K. Darwish, Y. Samih, and A. Abdelali, "Arabic offensive language on Twitter: Analysis and experiments," 2021, *arXiv:2004.02192*.

[17] 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, doi: 10.1109/ACCESS.2020.2994950.

[18] M. E. Aragón, M. Á. Álvarez-Carmona, M. Montes-y-Gómez, H. J. Escalante, L. V. Pineda, and D. Moctezuma, "Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets," in *Proc. IberLEF@ SEPLN*, 2019, pp. 478–494. Accessed: Dec. 16, 2023. [Online]. Available: https://www.researchgate.net/profile/Mario-Aragon-2/publication/334973555_Overview_of_MEX-

[19] G. I. Sigurbergsson and L. Derczynski, "Offensive language and hate speech detection for Danish," 2023, *arXiv:1908.04531*.

[20] P. Fortuna, J. Rocha da Silva, J. Soler-Company, L. Wanner, and S. Nunes, "A hierarchically-labeled Portuguese hate speech dataset," in *Proc. 3rd Workshop Abusive Lang. Online*, Florence, Italy, 2019, pp. 94–104, doi: 10.18653/v1/w19-3510.

[21] T. Ranasinghe, I. Anuradha, K. Silva, H. Hettiarachchi, and M. Zampieri, "SOLD: Sinhala offensive language dataset," *Lang. Resour. Eval.*, pp. 1–41, 2024.

[22] Y. Jeong, J. Oh, J. Ahn, J. Lee, J. Moon, S. Park, and A. Oh, "KOLD: Korean offensive language dataset," 2022, *arXiv:2205.11315*.

[23] Ç. Çöltekin, "A corpus of Turkish offensive language on social media," in *Proc. 12th Conf. Lang. Resour. Eval. (LREC)*, 2020, pp. 6174–6184. [Online]. Available: https://aclanthology.org/2020.lrec-1.758/

[24] A. N. Acar and S. İ. Omurca, "Offensive language detection in social media for Turkish language," in *Proc. 8th Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2023, pp. 267–270. Accessed: Dec. 16, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10286812/

[25] Z. Pitenis, M. Zampieri, and T. Ranasinghe, "Offensive language identification in Greek," 2020, *arXiv:2003.07459*.

[26] E. Kebriaei, A. Homayouni, R. Faraji, A. Razavi, A. Shakery, H. Faili, and Y. Yaghoobzadeh, "Persian offensive language detection," *Mach. Learn.*, vol. 113, no. 7, pp. 4359–4379, Aug. 2023, doi: 10.1007/s10994-023-06370-5.

[27] T. S. Ataei, K. Darvishi, S. Javdan, A. Pourdabiri, B. Minaei-Bidgoli, and M. T. Pilehvar, "Pars-OFF: A benchmark for offensive language detection on Farsi social media," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 2787–2795, Oct. 2023, doi: 10.1109/TAFFC.2022.3219229.

[28] W. A. Laal, *The History of the Urdu Language*. Delhi, India: Mujtabai Press, 1920. [Online]. Available: https://www.loc.gov/item/2021666209/

[29] Zoya, S. Latif, F. Shafait, and R. Latif, "Analyzing LDA and NMF topic models for Urdu tweets via automatic labeling," *IEEE Access*, vol. 9, pp. 127531–127547, 2021, doi: 10.1109/ACCESS.2021.3112620.

[30] S. Hussain, M. S. I. Malik, and N. Masood, "Identification of offensive language in Urdu using semantic and embedding models," *PeerJ Comput. Sci.*, vol. 8, p. e1169, Dec. 2022, doi: 10.7717/peerj-cs.1169.

[31] M. Wiegand, J. Ruppenhofer, and T. Kleinbauer, "Detection of abusive language: The problem of biased datasets," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Minneapolis, MN, USA, 2019, pp. 602–608, doi: 10.18653/v1/N19-1060.

[32] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. AbdelMajeed, and T. Zia, "Abusive language detection from social media comments using conventional machine learning and deep learning approaches," *Multimedia Syst.*, vol. 28, no. 6, pp. 1925–1940, Dec. 2022, doi: 10.1007/s00530-021-00784-8.

[33] K. R. Talpur, S. S. Yuhaniz, and B. Ali, "Cyberbullying detection in Roman Urdu language using lexicon based approach," *J. Crit. Rev.*, vol. 7, no. 16, pp. 834–848, 2020.
[34] H. Rizwan, M. H. Shakeel, and A. Karim, "Hate-speech and offensive language detection in Roman Urdu," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 2512–2522, doi: 10.18653/v1/2020.emnlp-main.197.
[35] M. M. Khan, K. Shahzad, and M. K. Malik, "Hate speech detection in Roman Urdu," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 1, pp. 1–19, Jan. 2021, doi: 10.1145/3414524.
[36] T. Sajid, M. Hassan, M. Ali, and R. Gillani, "Roman Urdu multi-class offensive text detection using hybrid features and SVM," in *Proc. IEEE 23rd Int. Multitopic Conf. (INMIC)*, Bahawalpur, Pakistan, Nov. 2020, pp. 1–5, doi: 10.1109/INMIC50486.2020.9318069.
[37] A. Atif, A. Zafar, M. Wasim, T. Waheed, A. Ali, H. Ali, and Z. Shah, "Cyberbullying detection and abuser profile identification on social media for Roman Urdu," *IEEE Access*, vol. 12, pp. 123339–123351, 2024.
[38] F. Razi and N. Ejaz, "Multilingual detection of cyberbullying in mixed Urdu, Roman Urdu, and English social media conversations," *IEEE Access*, vol. 12, pp. 105201–105210, 2024.
[39] A. Ullah, K. U. Khan, A. Khan, S. T. Bakhsh, A. U. Rahman, S. Akbar, and B. Saqia, "Threatening language detection from Urdu data with deep sequential model," *PLoS ONE*, vol. 19, no. 6, Jun. 2024, Art. no. e0290915.
[40] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding abuse: A typology of abusive language detection subtasks," in *Proc. 1st Workshop Abusive Lang. Online*, Vancouver, BC, Canada, 2017, pp. 78–84, doi: 10.18653/v1/W17-3012.
[41] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov, "SOLID: A large-scale semi-supervised dataset for offensive language identification," 2021, *arXiv:2004.14454*.

SALAH UD DIN received the B.S. degree from Gomal University and the M.S. degree in computer science from COMSATS University Islamabad. He is currently pursuing the Ph.D. degree with the University of Peshawar. He is a Lecturer with the Department of Computer Science, COMSATS University Islamabad. His research interests include artificial intelligence, natural language processing, big data analytics, machine learning, and deep learning.

SHAH KHUSRO (Member, IEEE) received the M.Sc. degree (Hons.) from the Department of Computer Science, University of Peshawar, Peshawar, Pakistan, and the Ph.D. degree from the Institute of Software Technology and Interactive Systems, Vienna University of Technology, Vienna, Austria. He is currently a Professor of computer science with the Department of Computer Science, University of Peshawar. He regularly publishes his work in reputed international journals and conferences. His research interests include web semantics, web engineering, information retrieval, web mining, search engines, human–computer studies, and ICT accessibility. He is a member of IEICE and ACM.

FARMAN ALI KHAN received the master's degree in computer science from the University of Peshawar, Peshawar, Pakistan, and the Ph.D. degree from Vienna University of Technology, Vienna, Austria. He is currently a Tenured Associate Professor with the Department of Computer Science, COMSATS University Islamabad, Pakistan.

MUNIR AHMAD (Senior Member, IEEE) received the Master of Computer Science degree from the Virtual University of Pakistan, in 2018, and the Ph.D. degree in computer science from the School of Computer Science, National College of Business Administration and Economics, in 2023. He has spent several years in the industry. He is currently the Executive Director/the Head of the IT Department, United International Group, Lahore, Pakistan, as an Associate Professor, NCBA&E, Pakistan, and a Research Professor with Korea University, Republic of Korea. He has vast experience in data management and efficient utilization of resources at multinational organizations. He has conducted many research studies on sentiment analysis and the utilization of AI to predict various healthcare issues. His research interests include data mining, big data, and artificial intelligence.

OUALID ALI received the Bachelor of Science degree in software engineering from Al Ain University, in 2011, the Master of Science degree in information technology management from The British University in Dubai, in 2013, and the Ph.D. degree in information science and technology from Universiti Kebangsaan Malaysia, in 2023. He was associated with The University of Manchester and The University of Edinburgh.

TAHER M. GHAZAL (Senior Member, IEEE) received the Bachelor of Science degree in software engineering from Al Ain University, in 2011, the Master of Science degree in information technology management from The British University in Dubai, in 2013, and the Ph.D. degree in information science and technology from Universiti Kebangsaan Malaysia, in 2023.

He was associated with The University of Manchester and The University of Edinburgh. He is currently a Seasoned Academician with a comprehensive educational background. He possesses over a decade of multifaceted expertise; he has fulfilled various roles, including a Lecturer, an Instructor, a Tutor, a Researcher, a Teacher, an IT Support/Specialist Engineer, and a Business/Systems Analyst. He has contributed significantly across diverse departments, such as Engineering, Computer Science, and ICT, and the Head of STEM and Innovation. His professional engagements have extended to governmental and private educational institutions under the purview of KHDA, the Ministry of Education, and the Ministry of Higher Education and Scientific Research, United Arab Emirates. His scholarly pursuits encompass a wide array of interests, including the IoT, artificial intelligence, cybersecurity, information systems, software engineering, web development, building information modeling, quality of education, management, big data, quality of software, and project management. He is actively engaged in community service through his involvement in impactful projects and research endeavors.