Kurdish Studies
January 2024
Volume: 12, No: 01 pp. 5079-5094
ISSN: 2051-4883 (Print) | ISSN 2051-4891 (Online)
www.KurdishStudies.net

DOI: 10.53555/ks.v12i1.3249

# Framework of Hate Speech Identification for Formal and Informal Text Using Lexical Approach

**Husnain Saleem[1*], Muhammad Javed[1], Syed Muhammad Ali Haider[1], Hamid Masood Khan[1], Muhammad Ahmad Jan[1], Asad Ullah[1]**

[1]Institute of Computing and Information Technology (ICIT), Gomal University, D.I. Khan, K.P.K, Pakistan.

***Corresponding Author:** Husnain Saleem
* Institute of Computing and Information Technology (ICIT), Gomal University, D.I. Khan, K.P.K, Pakistan.
**Email:** husnain@gu.edu.pk, ORCID: https://orcid.org/0009-0001-7513-1086

Received: 1 December 2023 Accepted: 28 December 2023 Published: 15 January 2024

**Abstract**
Social media refers to digital platforms and online venues where individuals and organizations share dynamic content and broadcast information. Through this dynamic virtual environment, prominent social networking sites such as Facebook, Instagram, Twitter, and YouTube allow users to produce, share, and trade different types of multimedia material such as text, photographs, videos, and links. Sentiment analysis is a digital process for determining and categorizing the emotional tone of textual material on social networking sites such as messages, comments, or tweets. It is also observed that this problem is extremely significant in the field of Natural Language Processing (NLP). Hate speech or Toxic speech is described in this context as language comprising hostile attitudes, insulting statements, and destructive intents directed against a person or a group of individuals. In this study, we used a lexicon-based approach at the sentence level to detect toxic speech in bilingual text specially published in English (Formal) and Roman-Urdu (informal) text. Moreover, in this study, we concentrated on three areas in particularly; race, religion, and nationality. We extracted our dataset from Twitter via the Twitter API, comprised of 3,030 tweets, 1,010 of which are relevant to each of the aforementioned domains. The proposed Framework attained outstanding average accuracy for race, religion, and nationality domains of **92.52%**, **93.03%**, and **93.35%**, respectively.

**Keywords:** NLP, Hate Speech, Toxic Speech, Roman Urdu, Lexicon Based Approach.

## 1. Introduction

In the world of communication social media platforms provide spaces, for people and organizations to freely express their thoughts, feelings and personal stories. These platforms hold a wealth of written information reflecting the voices of the community (Howard et al., 2021). In the developing world sentiment examination has arisen as a remarkable field, in Natural Language Processing (NLP). SA term was first utilized by (Nasukawa & Yi, 2003) who removed the feelings (good and pessimistic) for a particular subject rather than the entire record while the assessment mining term was first utilized by (Dave et al., 2003) who proposed a classifier for removing suppositions and relegate them good and pessimistic extremity. This region uses calculations to break down the content used on web-based entertainment stages. It intends to recognize and comprehend the scope of feelings communicated, whether they be good, pessimistic, or impartial (Tripathi & Naganna, 2015). Organizations utilize opinion examination to screen the public's feelings, via web-based entertainment stages empowering them to acquire experiences, into how clients see their brands and items (Bonta et al., 2019). Economic specialists depend on feeling investigation to survey shopper sentiments distinguish market patterns and settle on informed business choices (Banerjee et al., 2021). Organizations use opinion investigation to monitor their standing making a move to address any bad criticism and guaranteeing they keep a positive picture (Ibrahim & Wang, 2019). Feeling examination assumes a part, in movements as it considers the investigation of popular assessment, towards competitors and significant issues giving important experiences to shape movement techniques (Kušen & Strembeck, 2018). Clinical specialists use opinion investigation to assess the criticism given by patients, which permits clinics to upgrade the nature of care (Denecke & Deng, 2015). Market members, such, as merchants and financial backers depend on feeling investigation to evaluate the common market opinion. This assists them with deciding and overseeing gambles actually (Eachempati et al., 2021). Opinion examination assumes a part, in assisting organizations with acquiring bits of knowledge from client criticism empowering them to upgrade their items and administrations by recognizing regions, for development (Birjali et al., 2021). During seasons of emergency organizations use feeling examination to check assessment and make a move to limit any mischief to their standing (Ragini et al., 2018). These applications show the versatility and significance of SA in current information driven society.

Online entertainment is a double-sided deal, considering both positive and negative associations. On the in addition to side, it advances overall connectedness, permitting individuals to trade data, support noble cause drives, and participate in significant discussions (Smith, 2016). Nonetheless, others utilize similar media to spread toxic substance, which can affect viciousness, bias, and disruptiveness. Toxic discourse is characterized as an intentional assault on a particular gathering inspired by parts of their personality (De Gibert et al., 2018). Toxic/Harmful discourse or content on the web is characterized as the utilization of harming, forceful, or unjustifiable words or content focused on people or gatherings in view of qualities like race, religion, culture, orientation, sexual direction, or other safeguarded highlights (Silva et al., 2016). Toxic discourse on computerized destinations has created stresses over its capacity to move genuine harm, propagate bias, and subvert the standards of consideration and regard in informal organizations (Barker & Jurasz, 2021). An enormous number of the 123 papers on Hate speech discourse and correspondence sciences are from the US and the UK. These two countries have 71 examinations, with 48 from the US and 23 from the UK (Paz et al., 2020). (Awan, 2016) utilized a blended procedure with a grounded hypothesis to observe hate speech discourse connected with Muslims or Islam on person-to-person communication, particularly on Facebook. From his work, he inferred that 494 occasions of online harmfulness were designated towards Muslims. (Törnberg & Törnberg, 2016) utilized subject displaying and talk examination to perceive harmful discourse on Swedish web discussion. From their examination, they reasoned that Muslims have been used as a model to show a homogenous out-bunch engaged in battle, viciousness, and risky characteristics coming from Islam. (Awan & Zempi, 2016) utilized Twitter to perceive harmful discourse by utilizing blended strategy information gathering methods with a grounded hypothesis and figured out that Friendliness to Muslims is a continuous practice in both the web and the genuine world. (Pitsilis et al., 2018b) fostered a system for recognizing hate speech discourse in tweets utilizing profound learning procedures, as well as distinguishing harmful discourse in tweets utilizing RNN, which groups them into bigots, hottest, and ordinary substances. For two dialects (Hindi and English), (Bohra et al., 2018) proposed a structure for recognizing toxic substances and made a corpus of online entertainment writing for Hindi and English.

In light of the huge distinction in the general number of letters in order, mapping Urdu characters to an English console is illogical. Because of these difficulties, most Urdu speakers utilize Roman-Urdu (RU) while utilizing web-based entertainment locales, a substitute type of language that utilizes English letters to create Urdu words (Daud et al., 2015). Eight central points of contention with RU text examination were uncovered in the latest review (Mehmood et al., 2020). It contains words with a few spellings, words with various tones, and unpredictable upper casing.

The quick development of correspondence is made conceivable by virtual entertainment sites which affect individuals' ways of life. The development of sites like Facebook, Instagram, Twitter, YouTube, and so on. has modified the scene of content circulation for the advancement of items, administrations, strategies, and various associations. In any case, it is significant to notice that these sites can likewise be utilized to scatter unbearable talk against associations, legislatures, religions, and races. To resolve this issue, we adopted a dictionary-based procedure for identifying hate speech substances in bilingual material at the sentence level, with attention to race, religion, and ethnicity. In our examination, we center around Roman Urdu with English sentences to recognize non-toxic and toxic discourse. Rest of the paper is comprised of succeeding sections which are Section 1: Literature Review, Section 2: Research Methodology, Section 3: Results, Section 4: Comparative Analysis and Section 5: Conclusion and Future Work.

## 2. Related Work

The manner in which we draw in with each other and share thoughts in the period of innovation has generally changed because of the online social stages. Electronic social stages have pervaded each part of our life, with a tremendous variety of destinations like YouTube, Instagram, Facebook, Twitter, Snapchat and LinkedIn promptly accessible (Boateng & Amankwaa, 2016). People groups are sharing considerations, feelings, information, assessments, and experiences communicating in words/text by utilizing virtual entertainment stages like Twitter, web journals audit sites, etc. It has significantly had an impact on the method of individuals to speak with one another. Without a doubt, the long-range informal communication sites permit everyone to impart their insights (Alessia et al., 2015). (Morinaga et al., 2002) proposed a model utilizing text mining methods that naturally bring assessment of individuals about a particular item. These days, OM is enchanting the consideration of analysts for dealing with NLP and text mining. The approach of computerized innovation which gives the accessibility of person-to-person communication sites to everybody, there has been expanded in imparting web insights, assessing sites, items motion pictures, and so forth. which has taken the awareness of state-run administrations, clients, and associations to investigate and look at these feelings (Saberi & Saad, 2017). (Sharma et al., 2018) proposed a model (electronic application) for examining the opinions of tweets which empowers the client to see the new pattern of feeling investigation with the assistance of a related catchphrase. (Thet et al., 2010) involved fined-grained examination for tracking down close to the home direction and strength of a commentator about the film. (Singh et al., 2013) proposed a system for working out opinion extremity of web journal posts and film surveys by utilizing SVM, Senti WordNet and naive bayes. (Yang et al., 2016) proposed a model to compute the feelings (mental and physiological) of client in wellbeing circle utilizing changed LDA named conLDA. (Yadav et al., 2020) proposed a structure in light of multinomial naive bayes, support vector classifier, Stochastic Slope Plummet, and LSTM for the assessment examination of Punjabi-English text on person-to-person connections locales. Because of long choice relational communication locales, it is more straightforward for the association and states to pursue choices in light of the fact that each individual has a place with various society and culture are imparting their considerations and insights. (Yan et al., 2014) worked on English-Chinese text for deciding assessment examination by utilizing N-gram, regular language models, SVM, and stop word rundown of the two dialects. To perceive abstract data in literary information and find out their semantic direction, feeling examination is utilized.. (Al-Rowaily et al., 2015) fostered a dictionary for cybercrime to dissect dim sites for 2 dialects, one is Arabic and other is English. SentiLEN and SentiLAR dictionaries were made for languages individually. For the expectation of word demeanor in two dialects English and Chinese (Liu et al., 2016) fostered a framework utilizing Deep Learning (DL) methods like word vector, recurrent neural network, guileless bayes, and LSTM. A typical event in informal communication locales made by multilingual clients is code-mixing which implies that their one sentence/tweet is made out of more than one language. (Bali et al., 2014) dealt with the Hindi-

English text characterization and recommended that there ought to be a programmed framework to deal with code-mixing. (Lee et al., 2013) used feeling assessment to remove the assessment contained in each thought and comment assembled from a site for building an idea structure, which can help firms with perceiving arranged considerations for their improvement among endless contemplations. SA generally centres around audits of items and motion pictures, individuals' perspective on schooling, governmental issues, finance religion and so on. We are restless about the distinguishing proof of toxic discourse in web-based person-to-person communication sites. Web toxic discourse is a specific sort of internet based text which center around mishandling people groups openly (Cohen-Almagor, 2011).

Recently, toxic substances have acquired ubiquity as a topic. This is exhibited by mutually the growing political spotlight on the subject as well as the developing television's dedication to it (Fortuna & Nunes, 2018). (Bohra et al., 2018) anticipated a system to distinguish toxic discourse in two dialects, one English and the other is Hindi. They likewise created a body with virtual entertainment words for Hindi and English dialects. (Ting et al., 2013) consolidates informal organization examination with message mining strategies to comprehend how different toxic gatherings on Facebook share their contemplations and draw in clients. (Pitsilis et al., 2018a) proposed a system for the recognizable proof of toxic discourse in tweets utilizing profound learning method and furthermore distinguishes harmful discourse in tweets by utilizing RNN what separates them into bigot, hottest and ordinary substance. (Defersha & Tune, 2021) utilized converse report, N-gram, AI methods and term recurrence to make a framework which distinguishes harmful and toxic discourse via virtual entertainment in Afan Oromo which mark the tweet into oppressive, toxic and clean. (Alfina et al., 2017) utilized various highlights like person and word n-gram, pessimistic opinions and four AI methods for the recognition of toxic discourse in Indonesian linguistic and presumed that n-grams are the finest element for expectation. (Del Vigna12 et al., 2017) introduced a model utilizing the DL model and ML model to distinguish toxic discourse and figured out that the presentation of the two models is comparative. (Kwok & Wang, 2013) distinguished the harmful discourse for dark people groups on Twitter utilizing a supervised ML approach. (Warner & Hirschberg, 2012) distinguished the harmful discourse based on sexual direction religion and orientation by utilizing a format-based approach.

**Table 2.1:** Literature Review of Hate speech/Toxic speech detection

| Study | Problem Discourse | Method | Dataset | Language | Limitations | Forthcoming Work |
|---|---|---|---|---|---|---|
| (Zhang et al., 2018) | Rise of digital toxic speech has received substantial attention from corporations and government. | DNN | Twitter | English | It can be hard to recognize the presence of abstract origins such as sexism and racism. | Detection of abstract origins. |
| (Jokić et al., 2021) | Framework that aids in the improvement of online communities by recognizing harmful speech. | Lexicon resources with ML. | Twitter | Serbian | Negation, emoticons, changes the meaning of content. | Enhancing the corpus AbCoSER with News comment. |
| (Ibrohim & Budi, 2019) | Recognition of toxic and harmful content on social networking sites. | ML Procedures with SVM, RFDT etc. | Twitter | Indonesian | Unable to identify levels, targets and classes of toxic tongue. | Detection of stages, aims and groups of toxic content. |
| (Corazza et al., 2020) | On many social media platforms, violent and toxic information is on the rise. | Neural Architecture with word embedding. | (Bosco et al., 2018; Waseem & Hovy, 2016) | English, German and Italian | Limited only to 3 languages. | Detection in other languages. |
| (Perifanos & Goutsos, 2021) | Detecting abusive and toxic speech is a difficult phenomenon. | Multimode approach and computer vision. | Twitter | Greek | Detects toxic speech in single tweet. | Combining multimode approach with social graph information. |
| (Wich et al., 2022) | There has been little study on telegraph toxic content. | BERT | Telegram | German | Only spots in German linguistic. | Use of NN architecture benefits in improving performance. |
| (Arofah, 2018) | Study of potentially toxic web content related to blasphemy allegations against Basuki Purnama. | Rhetoric model | Political opinion and news. | Indonesian | The Toxic tongue rhetoric ignores the ethos and logos aspects. | Model to detect ethos & logos features. |
| (Khan et al., 2021). | Toxic content is a contentious issue that must be isolated and managed. | LR | Twitter | Roman Urdu | Tweets needs class balancing. | Detection in other languages. |

| Study | Problem Discourse | Method | Dataset | Language | Limitations | Forthcoming Work |
|---|---|---|---|---|---|---|
| (Hettiarachchi et al., 2020) | Detection of toxic material on social networking media. | LR, RF, MNB, SVM | Facebook | Romanized Sinhala | Modification of algorithm phonetic is not a positive technique. | Algorithm for Sinhala linguistic (Romanized) is required. |
| (Davidson et al., 2017) | Unreliable findings since they designate any text that contains a certain term as toxic content. | Trigram, bigram & unigram | Dataset from Davidson et al. (2017) | English | For detection in English linguistic only. | Focus more intently on the social media which host toxic content. |
| (Al-Makhadmeh & Tolba, 2020) | On the internet, toxic content may be predicted automatically. | ML with NLP. | Twitter | English and Arabic | For detection in text only. | Detection in audio & video. |
| (Gröndahl et al., 2018) | Automated identification model for toxic speech. | LR with Char level features. | Dataset from Wikipedia. | English | Only for English language. | Comparison b/w linguistic characteristics of various forms of toxic speech. |
| (Vidgen & Yasseri, 2020) | On the internet, there is toxic anti-Islamic discourse. | SVM | Tweets of handlers which tails political parties of UK. | English | Enhancement is required because toxic speech changes. | Automatic model which identifies challenging task in regard of toxic speech. |
| (MacAvaney et al., 2019) | Toxic discourse is spreading alongside the expansion of digital data. | Supervised learning with SVM. | Storm front dataset | English | For detection in English linguistic only. | Detection in other languages. |
| (Shibly et al., 2021) | Toxic discourse and its contents must be identified. | Monkey Learn libraries. | Kaggle | English | No appropriate system to recognize. | ML technique is required for the improvement in toxic content identification. |
| (Abro et al., 2020) | Automatically recognition of the toxic content on many datasets. | SVM | Twitter | English | Incapable to identify in actual time. | Predict how severe toxic speech text would be. |
| (Miok et al., 2022) | Removal of abusive content which needs a reliable toxic speech detector. | Bayesian method using Monte Carlo. | Facebook reviews. | English | Proposed system does not show any improvement in predictive performance. | Appropriate ML technique can be used to transformer networks. |
| (Sreelakshmi et al., 2020) | Identification of harsh and illogical words directed at anybody in an unpleasant manner. | SVM | Twitter | English & Hindi | Only for English & Hindi Language. | Insult, rude, and vulgarity detection in toxic content. |
| (Plaza-Del-Arco et al., 2021) | Automatic identification of hostile and toxic internet material. | Transformer-based model. | Twitter | Spanish | Multitasking learning improvement is needed. | Sarcasm and irony detection in toxic speech. |
| (Waseem et al., 2018) | Current statistics only contain some types of toxic words, such as racism or sexism. | DL | Dataset from Waseem (2016) | English | For detection in English linguistic only. | The utilization of data for catching toxic speech among dataset. |
| (Pariyani et al., 2021) | Toxic discourse may be identified using NLP techniques such as CNN. | ML with NLP. | Twitter | English | Unable to predict correctly. | Individual characteristics and motivation of a user should be considered. |
| (Rodriguez et al., 2019) | Toxic speech and communications have been an issue since the beginning of the internet. | Unsupervised approach using K-Mean Clustering. | Facebook | English | For detection in English linguistic only. | Text and Image toxicity prediction. |
| (Jaki & De Smedt, 2019) | A framework that can identifies toxic speech by itself. | Unsupervised approach using K-Mean Clustering. | Twitter | German | For detection in German linguistic only. | To increase the reliability of our model by means of DL. |

| Study | Problem Discourse | Method | Dataset | Language | Limitations | Forthcoming Work |
|---|---|---|---|---|---|---|
| (Badjatiya et al., 2017) | Toxic word detection on Twitter is critical for apps such as controversial event. | Hybrid Technique with Random Embedding. | Dataset from (Waseem & Hovy, 2016). | Roman Urdu | Limited to one dataset | To investigate the worth of the task-related handler system characteristics. |
| (Alshalan & Al-Khalifa, 2020) | Few models for toxic tweets in Arabic. | Hybrid Technique with word2vec. | Twitter | Arabic | For detection in Arabic linguistic only. | To check the accuracy of the model on other datasets. |

(Davidson et al., 2017) in his examination speaks about lexical distinguishing recognition procedures often yield mistaken outcomes since sorting every correspondence which incorporate specific expressions as toxic discourse, and SL employed research is not able to separate toxic discourse and hostile discourse in the past. In this study, 33,458 tweets utilizing API that are separated into 3 gatherings utilizing publicly supporting, as indicated by whether they contain toxic discourse, simply questionable language, or not one or the other. To separate between these few classes, a multi-class classifier is trained. Little part of the data identified by the vocabulary were considered to incorporate toxic discourse by human annotators, showing the constraints of lexical methodologies for recognizing possibly unsafe expressions. Automatic classification calculations might recognize among these few groupings with a decent lot of accuracy, an intensive investigation of the information exhibits that the consideration or absence of disparaging or harmful texts can work with and convolute viable classification. (Zhang et al., 2018) suggested a structure for the identification of online toxic discourse by utilizing DNN, GRN and convolutional. Monitoring of toxic discourse massive venture is drawn from organizations and government. Twitter dataset was utilized and their results displays that distinguishing harmful discourse in light of sexism and bigotry is certainly not a simple assignment. Tentative arrangement to deal with various organization structures for practicing highlights. (Gröndahl et al., 2018) offered a recognition technique that is helpless against rivals who might add grammatical mistakes, modify word cutoff points, or attach innocuous terms to the really toxic discourse. They reproduce seven states of the art toxic discourse distinguishing proof frameworks from prior studies and show that it is just compelling after assessed on the very sorts of text. They utilized six attacks to all of the 7 model mixes that had been duplicated from before research. While assault adequacy changed between models and datasets, most of the attacks significantly decreased the proficiency of every one of the 7 toxic discourse identifiers. Future exploration ought to focus on the datasets as opposed to the models. There should be more examination contrasting the language attributes of different types of harmful discourse (bigot, orientation, individual attacks, and so on.), likewise the qualifications between harmful and hostile text. As per (Khan et al., 2021), much exploration has been finished for the distinguishing proof of harmful discourse in EU dialects. In any case, South Asian dialects with restricted reserves definitely stand out enough to be noticed, leaving a huge number of people defenseless on person-to-person communication destinations. To our ongoing information, no examinations on the subject of distinguishing toxic language created in Roman Urdu have been led. They physically handled more than 90,000 tweets to find 5,000 Roman Urdu tweets by scratching them by utilizing an iterative cycle to make models, which were then used to deliver the Toxic Discourse RU 2020 corpus. Three levels of order are utilized to bunch the texts in this corpus specifically hostile, nonpartisan and basic. (Jokić et al., 2021) recommended the first corpus for acknowledgment of oppressive substance in the Serbian Language named AbCoSER that aides in making person-to-person communication sites a superior spot. 6,436 tweets were utilized by applying Half breed approach (AI strategies and dictionaries). Moreover, it is difficult to identify harmful substance since utilization of emojis, mockery and incongruity changes the importance of the discourse. Likely arrangements to improve AbCoSER with News remarks. (MacAvaney et al., 2019) dissected and featured issues with online mechanized strategies for toxic discourse recognizable proof by utilizing a multi-view SVM calculation that is less convoluted and conveying more plainly perceived ends than brain draws near and the exhibition that is nearly at the bleeding edge. Dataset utilized were 24,802 tweets, 16,914 tweets which were named as not one or the other, bigot and hottest. The previous occasions and developing suppositions toward specific issues present a deterrent for independent harmful discourse location frameworks. The need of computerized toxic discourse location devices develops in light of the fact that harmful discourse is a social issue. They examined the current strategies for accomplishing this work as well as a spic and span strategy that gets good precision. Furthermore, they set forth a new system that can show improvement over current techniques. Future work lies in that there is an interest for more noteworthy concentrate on this issue, considering the two of them specialized as well as reasonable issues, considering every one of the impediments are as yet present. (Ibrohim & Budi, 2019) proposed a system for acknowledgment of harmful discourse, oppressive items and perceived the objectives, classifications and levels of toxic language by applying ML methods with SVM, NB, RFDT, BR, LP and CC to forestall the questions among people groups via web-based entertainment. The disadvantage of the examination was that they can't recognize harmful discourse levels, targets and classes in multi mark grouping and proposed word implanting procedure for that reason. (Corazza et al., 2020) assessed for online harmful substance ID in 3 dialects (Italian, German and English) and thought about 1800, 1224 and 1080 potential settings for Italian, German and English separately to control forceful and toxic substance ascending on person-to-person communication sites by applying brain engineering with word implanting and n-gram. Likewise made sense of the job of emoticons and hashtag in distinguishing want content (Perifanos & Goutsos, 2021) made a system for the discovery of harmful substance particularly xenophobic, bigot and scornful substance in Greek language tweets by utilizing multimode approach with NLP and PC vision however it just recognizes toxic discourse in single tweet. Their model is openly accessible which is prepared on Greek tweets. In ongoing they are wanting to join multimode approach with social chart data. (Waseem et al., 2018) states that current datasets just incorporate specific disdain discourse classes, similar to bigotry and sexism, or explicit socioeconomics, similar to Americans, which unfavorably affects review while sorting material that is excluded from the preparation occurrences and analyzed methods for defeating naming and information gathering for tweets with hostile language, for example, unique marking plans, labels, or provincial and social impact by utilizing Perform MLT move toward on ML model. On account of MLT, the information implanted in one example to more readily suit one more by sharing it across numerous objectives. Future work to zero in on the usage of segment factors like pay, orientation, and mature as well as client information as extra hints for spotting oppressive and harmful discourse among datasets. (Pariyani et al., 2021) utilized NLP methods to identify disdain addresses on web. For text order issues CNN strategies surpass existing systems for text characterization issue. For identifying toxic discourse NLP, for example, CNN was utilized on English tweets. It allots the tweet to one of the classifications of the twitter dataset (toxic, hostile language). The downside of their model was that the model erroneously perceives some non-toxic discourse as harmful discourse. The vast majority of the toxic class was messed up. Future objectives center around the single qualities and inspiration of a client. (Sreelakshmi et al., 2020) proposed a ML model to distinguish toxic discourse in Hindi and English in blended structure by taking 10k examples from various web-based entertainment destinations as harmful and non-harmful. Due to the surprising spelling and linguistic contrasts, electronic harmful discourse distinguishing proof faces a few troubles. This toxic data could show up in blended structure, which makes the cycle testing, especially for a country, for example, India having an enormous populace that communicates in numerous dialects. Their future work was to work on the recommended way to deal with classify toxic tweets as annoying, impolite, and dirty. (Hettiarachchi et al., 2020) detected the web-based harmful substance written in Sinhala (Romanized) language. For that reason, they utilized 4 ML calculations LR, RF, MNB, SVM on 2500 Facebook remarks and identified the toxic discourse on long range interpersonal communication sites and separated it from porn. They likewise changed the phonetic calculation for Sinhala yet it was anything but a lucky methodology. Their impending goal was to alter the phonetic calculation for the Sinhala (Romanized) language. As per (Abro et al., 2020), various issues have emerged because of online stages, especially the sharing and dispersal of harmful discourse. In their exploration they are looking at the accomplishment of 8 ML procedures (NB, SVM, KNN, DT, RF, LR, AdaBoost and MLP) and 3 element designing methodologies (TFIDF, Doc2vec and Word2vec) on a dataset clarified into 3 gatherings (Hostile, Disdain and Non-hostile) having 14509 tweets which is accessible openly and claims that their examination is extremely helpful to naturally recognize toxic messages. Restriction of their examination is that it can recognize toxic messages in 3 classifications and unfit to distinguish the level of brutality in messages which isn't sufficient to distinguish harmful discourse. Likely arrangements were to construct a model that can identify the level of cruelty in messages. (Javed et al., 2020) proposed a framework for bilingual text for the detection of sarcasm for Politics and product domains from formal (English) and informal (Roman Urdu) text, their experiments and results represented higher accuracy in comparison to baseline studies.(Al-Makhadmeh & Tolba, 2020) utilized NLP and ML procedure to recognize toxic discourse from interpersonal interaction sites as harmful discourse is one of the most ridiculously horrendous of this action and proposed a model that consequently ensures slight misfortune capability and greatest expectation exactness by utilizing twitter information. Their future objective was to apply NLP with ML ways to deal with identify harmful substance in two configurations (sound and video). As per (Rodriguez et al., 2019), Toxic discourse/messages has been an issue starting from the beginning of the web, but the ascent of web based systems administration locales has amplified it to unbelievable levels. To handle this issue (Rodriguez et al., 2019) created a model that can recognize harmful messages in subjects that are examined routinely on Facebook utilizing Facebook API to gather 1k remarks from Facebook most well-known pages which advances toxicness. Chart examination was utilized to check which pages are advancing toxicness. Future work to dissect the answers of the clients to specific posts to get the full story and mockery. (Vidgen & Yasseri, 2020) investigated Islamophobic toxic substances on long-range interpersonal communication sites by applying hyper boundaries. At first, they took 4,000 tweets from clients which follows UK ideological groups and prevailed with regards to making of a model which consequently recognizes the toxic substance into non-solid, and frail Islamophobic. Much work is required in making disparity among qualities and future work was to make a device that is consequently custom-made for the testing tasks. (Jaki & De Smedt, 2019) expresses that because of the fast ascent of hostile web-based discourse, there is a requirement for novel advancements which can perceive disdain/toxic discourse without help from anyone else to help people content observing.

This postures critical troubles, such as deciding exactly what comprises freedom of discourse as well as what is denied in a given country and figuring out the exact language qualities of disdain/toxic discourse. (Jaki & De Smedt, 2019) dissected the toxic talks on twitter be taking 50k harmful tweets that happened during the German races that occurred between Aug-2017 to Apr-2018 by utilizing techniques (Subjective and Quantitative). Not all tweets were toxic tweets, some among them was hostile tweets and some are unlawful tweets as per their regulation. (Shibly et al., 2021) proposed a system to distinguish irritating talks through AI. By utilizing monkey learn AI libraries which are consolidated with python program they established eight kinds of disdain discourse in particular sexual, nationality, societal position, sex, handicap, religion, age and orientation from Kaggle dataset. Be that as it may, they couldn't identify and control disdain discourses naturally.

They recommended that controlling disdain discourses utilizing AI will be appropriate. (Badjatiya et al., 2017) referenced in their review that harmful discourse recognizable proof on Twitter is fundamental for applications like petulant occasion assortment, making chatterbots utilizing man-made brainpower, point idea, and examination of feeling. (Badjatiya et al., 2017) utilized DL way to deal 16k marked tweets into 3 classes (Not one or the other, Bigot and Hottest) and found that the DL approach is superior to word n-gram strategies. In later they are wanting to examine the worth of the errand related client framework qualities. (Miok et al., 2022) worked on toxic discourse indicator to eliminate inconsiderate/offending items, boycott clients and content producers. Unwavering quality isn't accomplished by utilizing DNN which depend on engineering. Three distinct datasets were utilized for identifying can't stand discourse: English tweets, Croatian news remark and Slovene Facebook remarks.

Unwavering quality data accomplished by their model is superior to BERT. In any case, their outcomes have no improvement for prescient execution. Future works means to adjust other Bayesian methodologies, for example, Loot and transformer organizations. (Alshalan & Al-Khalifa, 2020) referred to in his review that a ton of work has been led to make robotized strategies for perceiving toxic substance as a response to the peculiarities of harmful texts on Twitter utilizing straightforward ML procedures and DNN approaches yet at the same time there is little exploration on identifying harmful discourse with respect to Arabic language. In their exploration, (Alshalan & Al-Khalifa, 2020) investigated different NN Models to distinguish the toxic discourse in tweets written in Arabic language.

For that reason, they made a new dataset that contains 9316 named (Toxic, Ordinary and Harmful) toxic tweets and observed that CNN is best for identifying toxic discourse in Arabic language. At long last, they recommended that it tends to be supported assuming dataset is multi-marked and it can perceive additional composing examples and subjects. (Plaza-Del-Arco et al., 2021) performed multiple tasks approach utilizing transformer-based model to make electronic method for recognizing vicious and disdainful discourse online by utilizing dataset of Spanish tweets, the presentation accomplished by their model demonstrates that opinion characterization and extremity is the way to assist MTL with displaying to accurately distinguish harmful discourse more.

Notwithstanding, constraints lie in the way that computational expense is higher for the arrangement of different corpora. Likely arrangement to deal with a perplexing model that can recognize mockery in tweets that could be critical for harmful discourse recognition. (Arofah, 2018) says that toxic discourse generally contains three components ethos, tenderness, and logos and made sense of the disdain talks of Basuki Tjahaya Purnama that hauls him into prison.

As per ethos component, the toxic discourse is overlooked the ethos thing which gives the dependability of the source, as per tenderness, the author is choosing words to hasten awful opinions and outrage from its objective market, and as indicated by logos, the vast majority of the abhorrence content draws its readers. Given the discoveries, it is basic to decisively treat harmful discourse. Thus, the proposed examinations have featured the investigation of known techniques for spotting harmful or noxious language and has incorporated an essential stage committed to working on the viability of toxic discourse acknowledgment inside the suggested structure.

## 3. Methodology

We suggested an unsupervised lexicon-based framework at sentence level to deal with the challenge of toxic speech identification for analysis over twitter. Despite the availability of several machine learning algorithms for toxic content detection, they lack an adequate categorization and detection of roman language phrases due to a lack of resources. This paper presents a framework for detecting and categorizing toxic tweets. The primary goal of this system is to make it easier for the community to use social media especially twitter in order to improve their goods, organizations, regulations, and even trends. ur framework is depicted in Fig.1 as having four major stages; Data Collection, Data Pruning, Toxic Speech Identification and Scoring Module.

### 3.1 Data Collection

Datasets for experimental purposes were acquired from twitter using the twitter APIs. Figure 3.1 depicts the research's two core areas, namely Religion, Race and Nationality, which were chosen to examine the efficacy of the suggested framework. The gathered tweets for each of these areas has been carefully preserved in separate files to allow for further processing and analysis.

### 3.2 Data Pruning

Text pruning is the method of cleaning and eliminating unnecessary data from a large dataset obtained during the extraction phase. Because of the inclusion of several unnecessary and undesirable tags that do not add to the analysis process, it is clear that the mined text cannot be used directly for experimentation.

The data obtained frequently contains noise in the form of URLs, tags, and links, which might be unnecessary or disturbing to the study. Data preprocessing is a typical approach in text mining to remove noise from retrieved text, making it more legible and coherent.

#### 3.2.1 Conversion of Upper Case into Lower Case

It is the initial stage of text standardization in which upper case letters are transformed to lower case letters using NLTK and Python. It maintains uniformity in text analysis by treating all words identically, regardless of capitalization. This uniformity is critical in preventing errors in word processing. Converting text to lowercase also improves text matching and comparison procedures, making it easier to discover particular words expressions in text data. Lowercasing also minimizes the amount of the vocabulary, which can result in more efficient memory utilization and speedier processing.

#### 3.2.2 Removal of Irrelevant Symbols

In SA, the removal of unnecessary symbols is a critical preprocessing phase in which non-essential letters, symbols, and punctuation marks are carefully removed from the text input prior to SA (Javed & Kamal, 2018). It improves text clarity by removing distracting components such as special characters, punctuation, and emoticons, resulting in better text reading and enhances stability in SA by guaranteeing that identical material is consistently evaluated, hence increasing the dependability of SA results.

### 3.2.3 Stop Word Removal

It is clear that not all words in a phrase serve the objective of communicating thoughts or feelings. In actuality, a large majority of the words in a phrase are irrelevant at any level of the categorization process, especially those that appear to build a sentence regularly.

Among them, "stop words" are quite common, and their importance in the sentiment categorization technique is minor. The elimination of stop words is a key text preprocessing technique used in NLP and SA. It comprises removing frequent terms, known as "stop words," from text data prior to in-depth analysis.

### 3.2.4 Tokenization

Tokenization is a critical text preparation step in the realm of NLP. It is critical in turning a sentence into discrete units known as tokens.

These tokens form the basis for a wide range of NLP activities. Tokenization is required for a number of reasons. It divides material into digestible chunks, assisting in text comprehension and analysis. These tokens serve as the foundation for extraction of features in NLP models.

### 3.2.5 Part of Speech Tagging

This stage gives grammatical tags to each token in the extracted text or tweet, leading in a linguistically organized representation.

Part of speech (PoS) tags classify words according on their grammatical functions, such as verbs, adverbs, adjectives, or nouns, and provide information about the language used. PoS tagging is used in linguistic analysis to highlight the linguistic properties of tweets.

### 3.2.6 Lemmatization

Lemmatization, a key text preparation technique in the field of NLP, transforms words into their standard or dictionary form, known as the "lemma." This lemma reflects a word's root or base form, and lemmatization is essential for numerous strong reasons. It normalizes word by simplifying affected words to their core forms, assuring uniformity in text data processing.

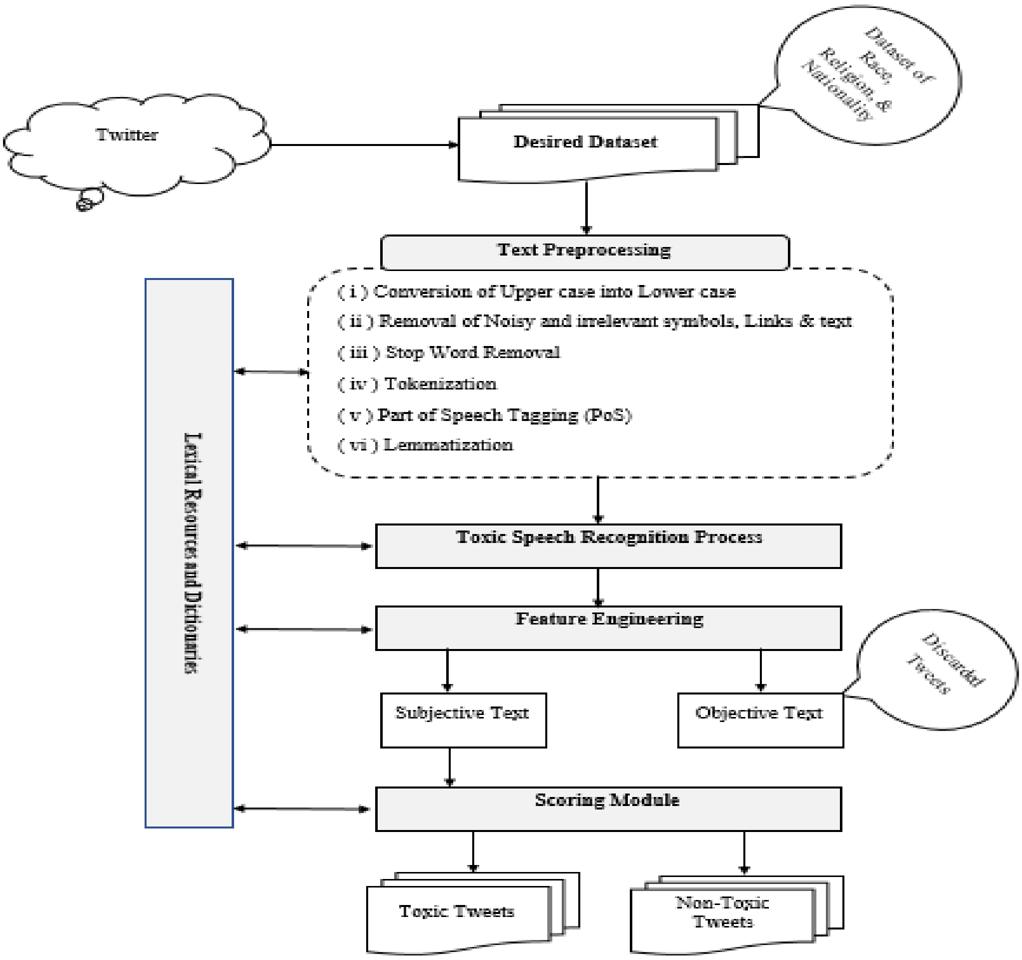
Twitter
Desired Dataset
Dataset of Race, Religion, & Nationality
Text Preprocessing
( i ) Conversion of Upper case into Lower case
( ii ) Removal of Noisy and irrelevant symbols, Links & text
( iii ) Stop Word Removal
( iv ) Tokenization
( v ) Part of Speech Tagging (PoS)
( vi ) Lemmatization
Lexical Resources and Dictionaries
Toxic Speech Recognition Process
Feature Engineering
Subjective Text
Objective Text
Discarded Tweets
Scoring Module
Toxic Tweets
Non-Toxic Tweets

**Fig 3.1:** Framework for toxic speech detection

### 3.3 Hate/ Toxic Speech Identification

Toxic speech is defined as targeting one or more groups of people based on their race, national origin, colour, or handicap, among other things. Most users on the internet used harsh language to mock someone based on hostility toward a person or a number of individuals during contact. In this study, we proposed a lexicon-based technique for identifying toxic speech for religion, nationality and race. The feature engineering method is employed for toxic speech detection, as demonstrated in Fig 3.1. When dealing with unsupervised machine learning algorithms, feature engineering performs a crucial role, and these characteristics have a direct influence on the correctness of the suggested framework. The annotated text, such as verbs, adverbs, and adjectives, is regarded as a feature, and these characteristics are regarded as subjective and objective hints. Tweets or words are characterized as subjective or objective based on these characteristics. In this research, we only looked at subjective tweets with toxic and non-toxic phrasing. The objective tweets are eliminated since they are useless for analysis.

### 3.4 Scoring Module

It is the final stage of the suggested framework. Senti Word Net is used to allocate weight to every opinionative token, If score > 0: Non-Toxic; If score < 0: Toxic; If score = 0: Neutral

## 4. Results

In this section the results of the probing strategy are presented thorough our proposed framework. A raw dataset of 3030 tweets in English, Roman Urdu, & English-Roman Urdu is utilized to evaluate the performance of Framework in three areas (Religion, Race, Nationality). A balanced dataset was created for each domain, with 1010 tweets for religion, 1010 tweets for race, and 1010 tweets for nationality remaining.
The structure of tweets is shown in Table 4.1.

**Table 4.1:** Statistics of tweets for Religion, Race and Nationality in English(Formal) and Roman-Urdu(Informal) text

| Domains | English Language (EN) | | Roman Urdu (RU) | | English-Roman Urdu (EN-RU) | | Total |
|---|---|---|---|---|---|---|---|
| | Non-Toxic | Toxic | Non-Toxic | Toxic | Non-Toxic | Toxic | |
| **1.Religion** | 250 | 250 | 125 | 125 | 130 | 130 | 1010 |
| **2.Race** | 150 | 175 | 250 | 300 | 60 | 75 | 1010 |
| **3.Nationality** | 300 | 200 | 310 | 90 | 50 | 60 | 1010 |
| **Total** | **700** | **625** | **685** | **515** | **240** | **265** | **3030** |

### 4.2 Precision

In ML, a model's accuracy measure called precision is used to quantify how effectively the framework predicts positive. Precision is an expression of the accuracy of a generated model employing machine learning in grouping positive samples. A good accuracy score indicates that the framework properly classifies positive data more frequently than it wrongly classifies positive data.

Precision is computed using the following formula:

$$\boldsymbol{P} = \frac{TP}{TP+FP} \qquad Eq. 4.2$$

### 4.3 Recall

Recall is a performance statistic in machine learning that measures how effectively a model can detect positive samples.

It calculates the percentage of true positives that the framework correctly anticipated. If the number of positive samples detected is bigger, the recall will be greater as well. In machine learning, recall is an important metric, especially when it comes to detecting all positive data.

The following formula is used to calculate recall:

$$\boldsymbol{R} = \frac{TP}{TP+FN} \qquad Eq. 4.3$$

### 4.4 F-Measure

In machine learning, the F-measure, also known as the F1 score, is used to quantify the accuracy of a two-dimensional classification model.

It is calculated by averaging memory and accuracy and assigning equal weight to each. When recall and accuracy are both important but one demands a bit more attention than the other, such as when FN is more important than FP or vice versa. The formula for calculating the F-measure is:

$$\boldsymbol{F1 - measure} = \frac{2(precision \times recall)}{(precision+recall)} \qquad Eq. 4.4$$

### 4.5 Accuracy

In machine learning, accuracy is a performance measure that is used to analyze how effectively a model predicts. It computes the model's accuracy by dividing correct predictions by total predictions.

Because it provides an easy-to-understand metric of a model's success, ML practitioners typically utilize accuracy as a statistic. The following formula is used to calculate accuracy:

$$\boldsymbol{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad Eq. 4.5$$

As indicated in Table 4.1, we examined and assessed a toxic speech detection framework for religion, race, and nationality tweets written in English, Roman Urdu, and a mix of Roman Urdu and English. English, Roman Urdu, and blend statistics Tables 4.3, 4.4, and 4.5 exhibit Roman Urdu-English tweets by religion, race, and nationality.

**Table 4.3:** Tweets extracted for the religion domain in Formal (Eng) and Informal (RU)

| True Positive | | | Total | False Positive | | | Total | Grand Total |
|---|---|---|---|---|---|---|---|---|
| **EN** | **RU** | **EN-RU** | | **EN** | **RU** | **EN-RU** | | |
| 210 | 110 | 125 | **445** | 15 | 8 | 8 | **31** | **476** |
| **True Negative** | | | **Total** | **False Negative** | | | **Total** | **Grand Total** |
| **EN** | **RU** | **EN-RU** | | **EN** | **RU** | **EN-RU** | | |
| 300 | 100 | 100 | **500** | 13 | 10 | 11 | **34** | **534** |

Where **EN**, **RU** and **EN-RU** represents English, Roman Urdu and English-Roman Urdu Language

**Table 4.4:** Tweets extracted for race domain in Formal (Eng) and Informal (RU)

| True Positive | | | Total | False Positive | | | Total | Grand Total |
|---|---|---|---|---|---|---|---|---|
| **EN** | **RU** | **EN-RU** | | **EN** | **RU** | **EN-RU** | | |
| 90 | 200 | 100 | **390** | 6 | 10 | 7 | **23** | **413** |
| **True Negative** | | | **Total** | **False Negative** | | | **Total** | **Grand Total** |
| **EN** | **RU** | **EN-RU** | | **EN** | **RU** | **EN-RU** | | |
| 324 | 119 | 106 | **549** | 15 | 19 | 14 | **48** | **597** |

**Table 4.5:** Tweets extracted for nationality domain in Formal (Eng) and Informal (RU)

| True Positive | | | Total | False Positive | | | Total | Grand Total |
|---|---|---|---|---|---|---|---|---|
| **EN** | **RU** | **EN-RU** | | **EN** | **RU** | **EN-RU** | | |
| 193 | 152 | 81 | **426** | 13 | 11 | 7 | **31** | **457** |
| **True Negative** | | | **Total** | **False Negative** | | | **Total** | **Grand Total** |
| **EN** | **RU** | **EN-RU** | | **EN** | **RU** | **EN-RU** | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 223 | 154 | 141 | **518** | 15 | 9 | 11 | **35** | **553** |

Now Precision of non-toxic and toxic tweets related to religion, race and nationality domains are shown in the table below in all languages English, Roman Urdu and English-Roman Urdu respectively.

**Table 4.6:** Precision of Religion, Race and Nationality

| **Domains Name** | **EN** | | **RU** | | **EN-RU** | |
|---|---|---|---|---|---|---|
| | **Non-Toxic** | **Toxic** | **Non-Toxic** | **Toxic** | **Non-Toxic** | **Toxic** |
| **Religion** | 93.33% | 95.84% | 93.22% | 90.90% | 93.98% | 90.09% |
| **Race** | 93.75% | 95.57% | 95.23% | 86.23% | 93.45% | 88.33% |
| **Nationality** | 93.68% | 93.69% | 93.25% | 94.47% | 92.04% | 92.76% |

Recall of non-toxic and toxic tweets related to religion, race, and nationality domains are shown in the table below in all languages English, Roman Urdu, and English-Roman Urdu respectively.

**Table 4.7:** Recall of Religion, Race and Nationality

| **Domains Name** | **EN** | | **RU** | | **EN-RU** | |
|---|---|---|---|---|---|---|
| | **Non-Toxic** | **Toxic** | **Non-Toxic** | **Toxic** | **Non-Toxic** | **Toxic** |
| **Religion** | 94.17% | 95.23% | 91.66% | 92.59% | 91.91% | 92.59% |
| **Race** | 85.71% | 98.18% | 91.32% | 92.24% | 87.71% | 93.80% |
| **Nationality** | 92.78% | 94.49% | 94.4% | 93.33% | 88.04% | 95.27% |

F1-Measure of non-toxic and toxic tweets related to religion, race, and nationality domains are shown in the table below in all languages English, Roman Urdu, and English-Roman Urdu respectively.

**Table 4.8:** F1-Measure of Religion, Race and Nationality

| **Domains Name** | **EN** | | **RU** | | **EN-RU** | |
|---|---|---|---|---|---|---|
| | **Non-Toxic** | **Toxic** | **Non-Toxic** | **Toxic** | **Non-Toxic** | **Toxic** |
| **Religion** | 93.74% | 95.53% | 92.43% | 91.73% | 92.93% | 91.32% |
| **Race** | 89.54% | 96.85% | 89.54% | 96.85% | 90.48% | 90.98% |
| **Nationality** | 91.66% | 94.08% | 93.82% | 93.89% | 89.99% | 93.99% |

Fig 4.1 shows the graphical representation of precision, recall, and f1-measure for the non-toxic and toxic speech of religion domain respectively.

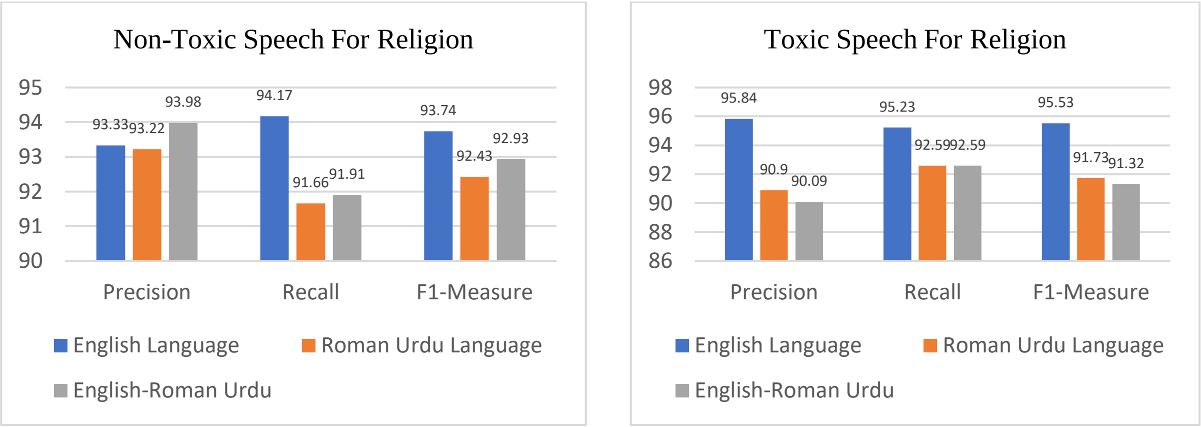

**Fig 4.1:** Precision, Recall and F1-Measure for non-toxic and toxic speech of religion domain

Fig 4.2 shows the graphical representation of precision, recall and f1-measure for non-toxic and toxic speech of race domain respectively.

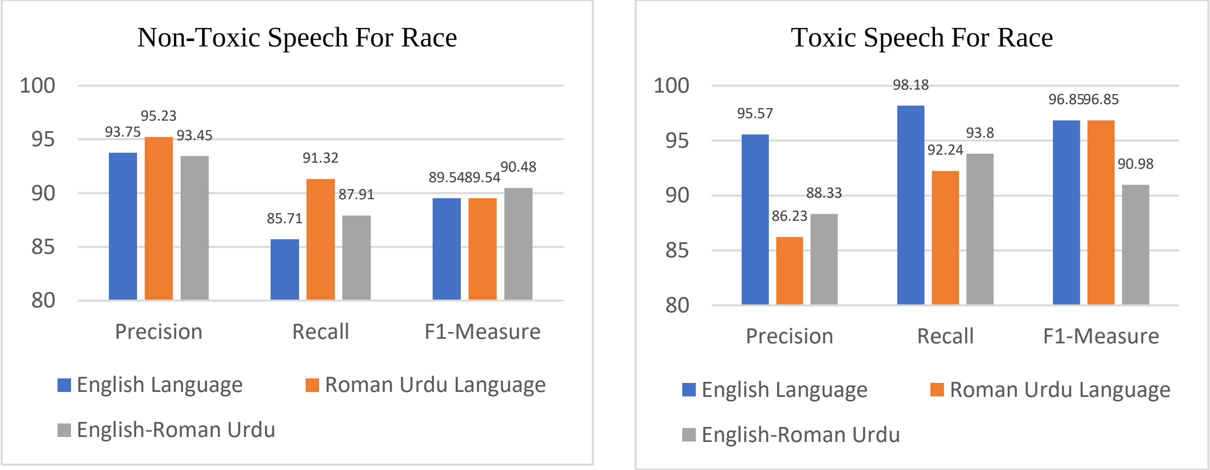

**Fig 4.2:** Precision, Recall and F1-Measure for non-toxic and toxic speech of race domain

Fig 4.3 shows the graphical representation of precision, recall and f1-measure for non-toxic and toxic speech of nationality domain respectively.

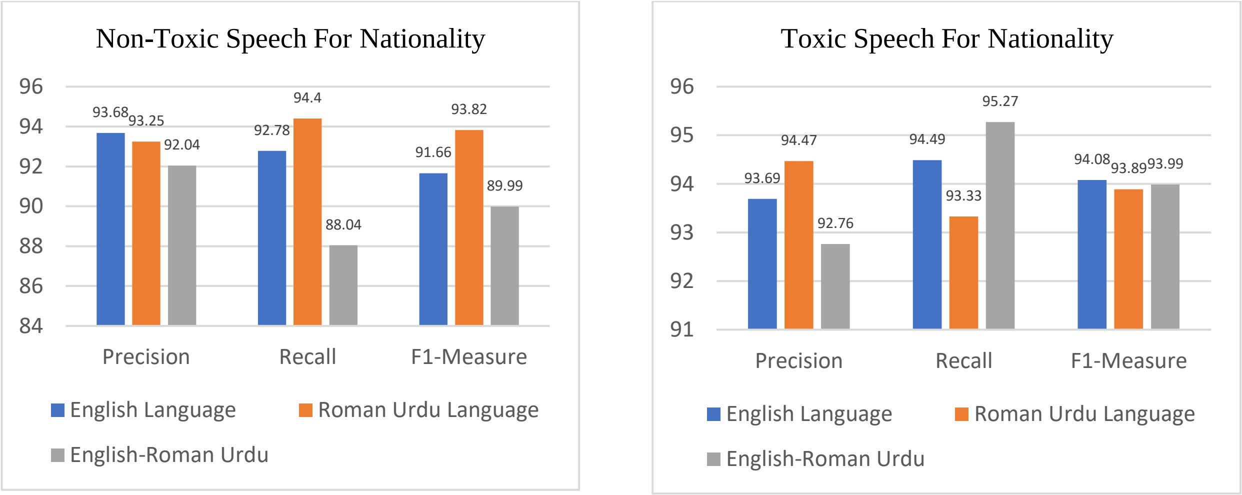

**Fig 4.3:** Precision, Recall, and F1-Measure for the non-toxic and toxic speech of the nationality domain

Accuracy of non-toxic and toxic tweets related to religion, race, and nationality domains are shown in the table below in all languages English, Roman Urdu, and English-Roman Urdu respectively.

**Table 4.9:** Accuracy of Religion, Race and Nationality domains

| Domains Name | Accuracy | | |
|---|---|---|---|
| | EN | RU | EN-RU |
| Religion | 94.79% | 92.1% | 92.21% |
| Race | 95.17% | 91.66% | 90.74% |
| Nationality | 93.69% | 93.86% | 92.5% |

Fig 4.4 shows the graphical representation of accuracy for all domains.

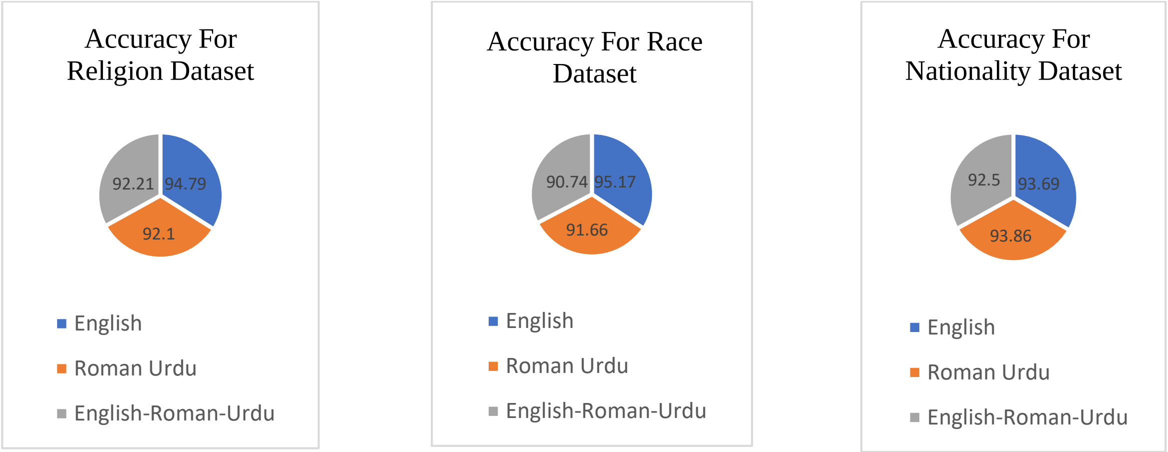

**Fig 4.4:** Accuracy for Religion, Race, & Nationality with domains regarding *w.r.t* English (Formal) & Roman Urdu (Informal).

## 5. Comparative Analysis

This study conducted a relative analysis of the suggested strategy in relation to prior research, taking into account the goal problem, accessible datasets, and analytical framework. The results show that our framework outperforms competing Naïve Bays, multi-layer perception, etc. when dealing with toxic speech recognition. As shown in Table 5.1, 5.2, and 5.3 our Proposed framework outperforms the existing studies.

**Table 5.1:** Comparative analysis of the Religion domain

| Study | Approach | Accuracy |
|---|---|---|
| (Ibrohim & Budi, 2019) | RFDT | 77.36 % |
| (Warner & Hirschberg, 2012) | SVM | 94 % |
| **Proposed** | **Lexicon Based** | **94.79 %** |

**Table 5.2:** Comparative analysis of Race domain

| Study | Approach | Accuracy |
|---|---|---|
| (Şahi et al., 2018) | NB | 75 % |
| (Gaydhani et al., 2018) | TFIDF | 95.6 % |
| **Proposed** | **Lexicon Based** | **95.17 %** |

**Table 5.3:** Comparative analysis of Nationality domain

| Study | Approach | Accuracy |
|---|---|---|
| (Nugroho et al., 2019) | RFM | 72.2 % |
| (Putri et al., 2020) | MLP | 83.4 % |
| **Proposed** | **Lexicon Based** | **93.86 %** |

## 6. Conclusion and Future Work

Social media sites permit individuals to speak with each other through web-based gatherings and organizations. Many toxic content studies have been carried out, as shown in Tables 5.1, 5.2, and 5.3. Our suggested focus on recognizing toxic sentiment in English, Roman Urdu, and English-Roman Urdu tweets in three categories with average accuracy of 93.03%, 92.52%, and 93.35%, respectively.

This examination is restricted to English and Roman Urdu language. Just these tongues were involved in the research. Subsequently, the ends and thoughts examined in this study probably won't be appropriate for foreign languages. The examination's worth inside the limits of English and Roman Urdu isn't impacted by the limitations, be that as it may, as it actually adds critical data and experiences to the area of SA for those specific dialects. The ill-advised treatment of negation, nonetheless, was an unmistakable deficiency that arose during the excursion of this study. Present day calculations and very much kept-up with datasets were utilized, however, the hardships of negation had all the earmarks of being a significant test. Negation went unknown and ignored, which brought about misconceptions of sentiment. The misclassification of feelings decreased the understanding of knowing about members' genuine feelings and bargained the legitimacy of the research. Certain the significance of this limitation, upcoming examinations on SA activities ought to focus on further developing strategies for dealing with negation, opening the entryway for additional exact and modern assessments of sentiments in NLP. To fully comprehend the underlying meaning and emotional breadth of speech, one must first understand the context in which that dictates how thoughts are articulated. The accuracy and depth of the SA were hampered by a lack of context consideration, which may have an impact on the study's validity and reliability.

In the future, we will expand our research into recognizing harmful speech in foreign languages such as Hindi, Persian, Arabic, and so on. The precise handling of negation and context is a crucial uniqueness for comprehending the emotion of a client's tweet, which will be the focus of future investigations. We strongly encourage future academics to pursue careers in the fields of emotion analysis, evaluation mining, and hazardous discourse differentiating evidence.

**Disclosure**

This research is part of PhD. program enrolled by the primary & corresponding author at the Institute of Computing & Information Technology (ICIT), Gomal University, Dera Ismail Khan, Pakistan.

**References**

1. Abro, S., Shaikh, S., Khand, Z. H., Zafar, A., Khan, S., & Mujtaba, G. (2020). Automatic hate speech detection using machine learning: A comparative study. International Journal of Advanced Computer Science and Applications, 11(8).
2. Al-Makhadmeh, Z., & Tolba, A. (2020). Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. Computing, 102(2), 501-522.
3. Al-Rowaily, K., Abulaish, M., Haldar, N. A.-H., & Al-Rubaian, M. (2015). BiSAL–A bilingual sentiment analysis lexicon to analyze Dark Web forums for cyber security. Digital Investigation, 14, 53-62.
4. Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications, 125(3).
5. Alfina, I., Mulia, R., Fanany, M. I., & Ekanata, Y. (2017). Hate speech detection in the Indonesian language: A dataset and preliminary study. 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS),
6. Alshalan, R., & Al-Khalifa, H. (2020). A deep learning approach for automatic hate speech detection in the saudi twittersphere. Applied Sciences, 10(23), 8614.
7. Arofah, K. (2018). Rhetorical Analysis of Hate Speech: Case Study of Hate Speech Related to Ahok's Religion Blasphemy Case. Mediator: Jurnal Komunikasi, 11(1), 91-105.
8. Awan, I. (2016). Islamophobia on social media: A qualitative analysis of the facebook's walls of hate. International Journal of Cyber Criminology, 10(1), 1.
9. Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. Aggression and violent behavior, 27, 1-8.
10. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. Proceedings of the 26th international conference on World Wide Web companion,
11. Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014). "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. Proceedings of the first workshop on computational approaches to code switching,
12. Banerjee, S., Dellarocas, C., & Zervas, G. (2021). Interacting user-generated content technologies: How questions and answers affect consumer reviews. Journal of Marketing Research, 58(4), 742-761.
13. Barker, K., & Jurasz, O. (2021). Text-Based (Sexual) Abuse and Online Violence Against Women: Toward Law Reform? In The Emerald International Handbook of Technology-Facilitated Violence and Abuse (pp. 247-264). Emerald Publishing Limited.
14. Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems, 226, 107134.
15. Boateng, R., & Amankwaa, A. (2016). The impact of social media on student academic life in higher education. Global Journal of Human-Social Science, 16(4), 1-8.
16. Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). A dataset of Hindi-English code-mixed social media text for hate speech detection. Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media,
17. Bonta, V., Kumaresh, N., & Janardhan, N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. Asian Journal of Computer Science and Technology, 8(S2), 1-6.
18. Bosco, C., Felice, D. O., Poletto, F., Sanguinetti, M., & Maurizio, T. (2018). Overview of the evalita 2018 hate speech detection task. Ceur workshop proceedings,
19. Cohen-Almagor, R. (2011). Fighting hate and bigotry on the Internet. Policy & Internet, 3(3), 1-26.
20. Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A multilingual evaluation for online hate speech detection. ACM Transactions on Internet Technology (TOIT), 20(2), 1-22.
21. Daud, M., Khan, R., & Daud, A. (2015). Roman Urdu opinion mining system (RUOMiS). arXiv preprint arXiv:1501.01386.
22. Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. Proceedings of the 12th international conference on World Wide Web,
23. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of the international AAAI conference on web and social media,
24. De Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. arXiv preprint arXiv:1809.04444.
25. Defersha, N., & Tune, K. (2021). Detection of hate speech text in afan oromo social media using machine learning approach. Indian J Sci Technol, 14(31), 2567-2578.
26. Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. Proceedings of the First Italian Conference on Cybersecurity (ITASEC17),
27. Denecke, K., & Deng, Y. (2015). Sentiment analysis in medical settings: New opportunities and challenges. Artificial intelligence in medicine, 64(1), 17-27.
28. Eachempati, P., Srivastava, P. R., Kumar, A., Tan, K. H., & Gupta, S. (2021). Validating the impact of accounting disclosures on stock market: A deep neural network approach. Technological Forecasting and Social Change, 170, 120903.

29. Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4), 1-30.
30. Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. (2018). Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. arXiv preprint arXiv:1809.08651.
31. Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is" love" evading hate speech detection. Proceedings of the 11th ACM workshop on artificial intelligence and security,
32. Hettiarachchi, N., Weerasinghe, R., & Pushpanda, R. (2020). Detecting hate speech in social media articles in romanized sinhala. 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer),
33. Howard, P. N., Neudert, L.-M., Prakash, N., & Vosloo, S. (2021). Digital misinformation/disinformation and children. UNICEF. Retrieved on February, 20, 2021.
34. Ibrahim, N. F., & Wang, X. (2019). A text analytics approach for online retailing service improvement: Evidence from Twitter. Decision Support Systems, 121, 37-50.
35. Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. Proceedings of the third workshop on abusive language online,
36. Javed, M. and Kamal, (S., 2018). Normalization of unstructured and informal text in sentiment analysis. International Journal of Advanced Computer Science and Applications, 9(10).78-85.
37. Javed M, Ziauddin, Kamal S et al. (2020). Socio monitoring framework (SMF): Efficient sentiment analysis through informal and native terms. International Journal of Advanced and Applied Sciences, 7(12): 113-126.
38. Jaki, S., & De Smedt, T. (2019). Right-wing German hate speech on Twitter: Analysis and automatic detection. arXiv preprint arXiv:1910.07518.
39. Jokić, D., Stanković, R., Krstev, C., & Šandrih, B. (2021). A Twitter Corpus and lexicon for abusive speech detection in Serbian. 3rd Conference on Language, Data and Knowledge (LDK 2021),
40. Khan, M. M., Shahzad, K., & Malik, M. K. (2021). Hate speech detection in roman urdu. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 20(1), 1-19.
41. Kušen, E., & Strembeck, M. (2018). Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. Online Social Networks and Media, 5, 37-50.
42. Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. Twenty-seventh AAAI conference on artificial intelligence,
43. Lee, H., Choi, K., Yoo, D., Suh, Y., He, G., & Lee, S. (2013). The more the worse? Mining valuable ideas with sentiment analysis for idea recommendation.
44. Liu, G., Xu, X., Deng, B., Chen, S., & Li, L. (2016). A hybrid method for bilingual text sentiment classification based on deep learning. 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD),
45. MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. PloS one, 14(8), e0221152.
46. Mehmood, K., Essam, D., Shafi, K., & Malik, M. K. (2020). An unsupervised lexical normalization for Roman Hindi and Urdu sentiment analysis. Information Processing & Management, 57(6), 102368.
47. Miok, K., Škrlj, B., Zaharie, D., & Robnik-Šikonja, M. (2022). To ban or not to ban: Bayesian attention networks for reliable hate speech detection. Cognitive Computation, 14(1), 353-371.
48. Morinaga, S., Yamanishi, K., Tateishi, K., & Fukushima, T. (2002). Mining product reputations on the web. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining,
49. Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. Proceedings of the 2nd international conference on Knowledge capture,
50. Nugroho, K., Noersasongko, E., Fanani, A. Z., & Basuki, R. S. (2019). Improving random forest method to detect hatespeech and offensive word. 2019 International Conference on Information and Communications Technology (ICOIACT),
51. Pariyani, B., Shah, K., Shah, M., Vyas, T., & Degadwala, S. (2021). Hate speech detection in twitter using natural language processing. 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV),
52. Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate speech: A systematized review. Sage Open, 10(4), 2158244020973022.
53. Perifanos, K., & Goutsos, D. (2021). Multimodal hate speech detection in Greek social media. Multimodal Technologies and Interaction, 5(7), 34.
54. Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018a). Effective hate-speech detection in Twitter data using recurrent neural networks. Applied Intelligence, 48(12), 4730-4742.
55. Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018b). Effective hate-speech detection in Twitter data using recurrent neural networks. Applied Intelligence, 48, 4730-4742.
56. Plaza-Del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). A multi-task learning approach to hate speech detection leveraging sentiment analysis. IEEE Access, 9, 112478-112489.
57. Putri, T., Sriadhi, S., Sari, R., Rahmadani, R., & Hutahaean, H. (2020). A comparison of classification algorithms for hate speech detection. Iop conference series: Materials science and engineering,
58. Ragini, J. R., Anand, P. R., & Bhaskar, V. (2018). Big data analytics for disaster response and recovery through sentiment analysis. International Journal of Information Management, 42, 13-24.

59. Rodriguez, A., Argueta, C., & Chen, Y.-L. (2019). Automatic detection of hate speech on facebook using sentiment and emotion analysis. 2019 international conference on artificial intelligence in information and communication (ICAIIC),
60. Saberi, B., & Saad, S. (2017). Sentiment analysis or opinion mining: a review. Int. J. Adv. Sci. Eng. Inf. Technol, 7(5), 1660-1666.
61. Şahi, H., Kılıç, Y., & Sağlam, R. B. (2018). Automated detection of hate speech towards woman on Twitter. 2018 3rd international conference on computer science and engineering (UBMK),
62. Sharma, N., Pabreja, R., Yaqub, U., Atluri, V., Chun, S. A., & Vaidya, J. (2018). Web-based application for sentiment analysis of live tweets. Proceedings of the 19th Annual International Conference on Digital government research: Governance in the data Age,
63. Shibly, F., Sharma, U., & Naleer, H. (2021). Classifying and measuring hate speech in twitter using topic classifier of sentiment analysis. International Conference on Innovative Computing and Communications,
64. Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. Proceedings of the International AAAI Conference on Web and Social Media,
65. Singh, V., Piryani, R., Uddin, A., & Waila, P. (2013). Sentiment analysis of Movie reviews and Blog posts. 2013 3rd IEEE International Advance Computing Conference (IACC),
66. Smith, E. E. (2016). "A real double-edged sword:" Undergraduate perceptions of social media in their learning. Computers & Education, 103, 44-58.
67. Sreelakshmi, K., Premjith, B., & Soman, K. (2020). Detection of hate speech text in Hindi-English code-mixed data. Procedia Computer Science, 171, 737-744.
68. Thet, T. T., Na, J.-C., & Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of information science, 36(6), 823-848.
69. Ting, I.-H., Wang, S.-L., Chi, H.-M., & Wu, J.-S. (2013). Content matters: A study of hate groups detection based on social networks analysis and web mining. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,
70. Törnberg, A., & Törnberg, P. (2016). Muslims in social media discourse: Combining topic modeling and critical discourse analysis. Discourse, Context & Media, 13, 132-142.
71. Tripathi, G., & Naganna, S. (2015). Feature selection and classification approach for sentiment analysis. Machine Learning and Applications: An International Journal, 2(2), 1-16.
72. Vidgen, B., & Yasseri, T. (2020). Detecting weak and strong Islamophobic hate speech on social media. Journal of Information Technology & Politics, 17(1), 66-78.
73. Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. Proceedings of the second workshop on language in social media,
74. Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. Proceedings of the NAACL student research workshop,
75. Waseem, Z., Thorne, J., & Bingel, J. (2018). Bridging the gaps: Multi task learning for domain transfer of hate speech detection. Online harassment, 29-55.
76. Wich, M., Gorniak, A., Eder, T., Bartmann, D., Cakici, B. E., & Groh, G. (2022). Introducing an abusive language classification framework for telegram to investigate the german hater community. Proceedings of the International AAAI Conference on Web and Social Media,
77. Yadav, K., Lamba, A., Gupta, D., Gupta, A., Karmakar, P., & Saini, S. (2020). Bilingual sentiment analysis for a code-mixed Punjabi English social media text. 2020 5th International Conference on Computing, Communication and Security (ICCCS),
78. Yan, G., He, W., Shen, J., & Tang, C. (2014). A bilingual approach for conducting Chinese and English social media sentiment analysis. Computer Networks, 75, 491-503.
79. Yang, F.-C., Lee, A. J., & Kuo, S.-C. (2016). Mining health social media with sentiment analysis. Journal of medical systems, 40(11), 1-8.
80. Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on twitter using a convolution-gru based deep neural network. European semantic web conference,