ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

JOURNAL OF APPLIED LINGUISTICS AND TESOL

Vol.9. No.1.2026

AUTOMATIC IDENTIFICATION OF HATE SPEECH IN ONLINE COMMENTS: A TOOL-BASED LINGUISTIC ANALYSIS OF ROMAN URDU

Asma Batool
MPhil Scholar, Department of English, NUML University, Faisalabad Campus
Email: asmabatool556@gmail.com
Dr. Aftab Akram
Lecturer, Department of English, NUML University, Faisalabad Campus
Email: aakrum@numl.edu.pk

Abstract
This paper explores the automatic detection of hate speech in online comments by using a tool-based linguistic analysis of Pakistani Roman Urdu discourse in Twitter. Taking the exploratory model of computational design and a mixed-method approach, which is a combination of the automated rule-based comment detection and interpretative linguistic analysis, the research gathered a sample of 108 comments in absolute terms and addressed to Pakistani political leaders and institutions. A three-level hate speech lexicon was inductively built on the basis of the corpus and implemented with the help of the WordList and Collocation detection patterns of the AntConc corpus tool, to determine high-frequency markers of hate and its collocating patterns. The comments were then grouped as either hate speech, borderline, or non-hate speech. Critical Discourse Analysis and Speech Act Theory were the theoretical bases of the analysis. Results indicate that hate speech in this corpus is not a random or idiosyncratic event but a discursive practice that is structured and ideologically consistent, and that recreates the status quo hierarchies of gender, religion, ethnicity, political affiliation, etc., using language. More importantly, the analysis shows that a rule-based keyword detection, however productive as a first-pass tool, cannot be effectively used as an independent classification tool in Roman Urdu, 38.9% of comments are located in an ambiguous category, which needs pragmatic and contextual analysis in order to be resolved. This paper has its contribution in an approach to replicate a methodology, a curated Roman Urdu hate speech corpus, with an annotated corpus, and a sociolinguistically grounded lexicon to a field of digital discourse study that is underrepresented.

Keywords: *hate speech detection, Roman Urdu, AntConc, rule-based detection, Critical Discourse Analysis, Twitter.*

Introduction

The fast development of online communication has made online space one of the key places of public discussion. Although social media and news comment boards promote democratic engagement, they are now the places where hate speech propagates easily and at a high rate. The comments made by the online audience may support discrimination, legitimize aggression, and cause social polarization. This means that automatic hate speech recognition has attracted significant research in both computational linguistics and discourse studies. Conventional methods of analysis of hate speech tend to be based on either computational models or only a qualitative analysis of discourse. Nonetheless, the use of computational systems can miss out on the nuances of linguistic meanings, whereas manual discourse analysis is time-consuming and small-scale. Thus, the proposed research assumes an exploratory computational research design with an automated detection and an interpretive linguistic research. Through the incorporation of both computational and theoretical knowledge based on the Critical Discourse Analysis (CDA) and the

Speech Act Theory, the study will pose the question of not only understanding how hate speech can be automatically detected, but also the functionality of hate speech in online communication.

This study is mixed-method research. The online responses will be obtained via the selection of news websites or social media. The hate-speech keyword lexicon will be developed on the basis of the present literature and the observations of the context. The textual data will then be cleaned and normalized so as to have consistency in analysis. To analyze it automatically, the AntConc version 4.2.4 will be used as the major corpus software. The search for hate-related keywords, creation of word frequency lists, finding of concordance lines, and analysis of collocations will be conducted with the help of AntConc. It enables the research to identify repeated linguistic patterns in hate speech that do not compromise any interpretation.

In addition to automated classification, the research will involve an interpretive analysis of linguistic patterns causing hate labels. Based on the Critical Discourse Analysis, this study will examine the ways in which the language forms social identities, power relations, and ideologies. It will also analyze how hate comments can be used as an insult, threat, dehumanization, or exclusion using Speech Act Theory. This paper brings together corpus-based computational procedures and the linguistic theory in offering a critically informed tool-based study of hate speech on online commentary. The AntConc tool will also make the hate speech detection transparent, reproducible, and linguistically understandable, which perfectly fits the theoretical framework and research goals of the study.

Statement of the Problem

The evolution of internet sites has helped broadcast opinions and ideas of people faster, as well as hate speech, which has been encouraged to be transmitted easily. Internet remarks might comprise hate speech to people or social groups based on gender, religion, race, and other identity aspects, to encourage social disparities and propagate animosity. Even though increased computational methods of automatic detection are increasingly becoming available, a number of studies have not concentrated their attention on the linguistic regularities of nasty content or merely concentrate on machine learning frameworks. Repeat automated systems are based on black-box models which do not consider the contextual language or are based on false recognition of comments. It leaves a major gap; a keyword or rule-based detection of hate speech in linguistic approach, that is a combination of interpretive analysis and automated detection, is required in order to have the academics not merely know what hate speech is but how this is created and presented in a discourse online.

Significance of the Study

The importance of this research is that it addresses the gap in the knowledge and identification of hate speech over the Internet, both in the real world and in theory. It is a lexicon-based, rule-driven system deployed using the AntConc 4.3.1 tool that provides a transparent and predictable process of identifying hate speech in online comments that would help social media moderators, news agencies, and policy-makers to monitor and ban posts containing hate speech. The study also involves the use of the Critical Discourse Analysis and Speech Act Theory that allows more thoroughly analyzing the linguistic means and speech functions that help spread hate and, thus, lead to the understanding of how social identities, power relations, and ideologies are represented in digital discourse. The study methodology uncovers a mixed methodology, where the corpus-based computational analysis is combined with the qualitative interpretation, which can be taken as an example to follow in the linguistic, discourse analysis, and digital communication studies. Overall, the study introduces social and linguistic understanding of the phenomenon of hateful language on the internet through keyword-based detection, blurring the source of computational assistance and critical discourse analysis.

Research Questions

1. Which lexical patterns and keywords are most indicative of hate speech in online comments when analyzed through a rule-based approach in AntConc?
2. Which linguistic patterns, keywords, and collocations are most frequently associated with hate speech in online comments?
3. How do online comments enact hate speech through language, in terms of social meaning and speech acts, as interpreted through Critical Discourse Analysis and Speech Act Theory?

Research Objectives

1. To identify hate speech in online comments using a rule-based, lexicon-driven approach with AntConc.
2. To analyze the linguistic patterns, keywords, and collocations associated with hate speech.
3. To examine how hate speech functions socially and pragmatically using Critical Discourse Analysis and Speech Act Theory.

Literature Review

The automatic detection of emotions in text has always faced numerous challenges (Biagioni, 2016). It is followed by the problem of detecting offensive content in text, which is secondary to sentiment classification (Schmidt & Wiegand, 2019). Kwok and Wang in 2014 found that in the case of anti-black racism, the total of 86% of the tweets were found to be racist and discriminatory due to the presence of offensive and pejorative terms and language.

A couple of renowned competitions (e.g., SemEval-2019 [191] and 2020 [192] & GermEval-2018 [183]) have tried numerously to identify the method of automated hate speech detection by organizing different events. In this regard, Scholars have filled sizeable data with numerous datasets from various sources that stimulated research in the field. Many of these studies have also addressed hate speech in languages other than English and online communities. This motivated the exploration and comparison of different processing pipelines, such as the selection of feature set and Machine Learning (ML) (e.g., supervised, unsupervised, and semi-supervised), and classification (e.g., Naives Bayes, Logistic Regression (LR), Convolution Neural Network (CNN), LSTM, BERT deep learning architectures, and so on) (Jahan & Oussalah, 2023). In recent years, hate speech on social media has increased in various forms. This has severe results and social effects upon victims of every demographic (Mathew et al., 2021), to a great extent influences the state of their mental health (Saha et al.,2021), and even becomes a catalyst for hate crimes in the real world (Relia et al., 2019).

Consequently, automated detection of hate speech is especially significant, and research studies have been conducted numerous times in a bid to define and identify these hateful or toxic contents (Zhou et al., 2021; Lahnala et al.,2022; Kotarcic et al., 2022). Nonetheless, the existing hate speech detection techniques do not tend to be robust and generalizable in most cases. even to somewhat different data sets on the same task, which, to a great extent, does not allow their application in the practical world (Vidgen et al., 2019). This could be because the prevailing models are vulnerable to the spurious relationships amongst data and labels of training (e.g., hateful) (Ramponi and Tonelli, 2022), which could result in the bias of vulnerable and minority groups, such as African American Vernacular English speakers, and can perpetuate racism (Harris et al., 2022). The more frequent the detectors are, the more likely they are to be biased in case the identity words appear co-occurring with the hateful word in the training set (e.g., fine-tuned BERT) to commit false predictions during inference. Consequently, there is a dire need to fully conceptualize and establish the spurious hate speech detection correlations and more to mitigate the bias due to spurious correlations. An increasing number of recent studies have investigated the spurious hate speech-detecting correlation (Zhou et al., 2021; Kennedy et al., 2020; Sap et al.,2022) and discovered that many tokens that target minority groups are associated with the hate label (Bender et al., 2021). As a result, techniques such as concealing and eliminating tokens with the assistance of Human annotations have been made to reduce the spurious correlations (Ramponi and Tonelli, 2022).

In 2014, C.J. Hutto et al. suggested a strategy to classify the sentiment based on VADER, which is a rule-based system. Initially, they made a list of lexical features that are very sensitive to the sentiment of social media posts. Then they added that list of lexical features to five general rules, which summarize syntactical and grammatical rules to present sentiment intensity. Finally, they have discovered that VADER has completed 96% accuracy on the rule-based model on Twitter sentiments. A technique of determining the Sentiment Analysis of social media text via the Rule-based method was suggested by Dennis Gitari et al. in 2015. They also divided the hate speech issue in this work into three areas: religion, nationality, and race. The primary goal of this paper is to come up with a classification model that uses sentiment analysis. The model developed is not only able to identify subjective sentences but also classifies and ranks the polarity of sentiment phrases. Then, after this, they correlate the semantic and subjective characteristics with hate speech. Lastly, they obtained 71.55 percent accuracy with the lexicon method. Fatahillah et al. (2017) employed the Naive Bayes Classifier Algorithm to identify hate speech on Instagram with the use of the k-nearest neighbor classifier. They gathered the data set through the Twitter API and annotated the data set through manual means. They used the Naive Bayes Classifier algorithm after the preprocessing and feature engineering stage and discovered 93% of the accuracy.

M. Ali Fauzi et al. (2018) suggested a method of detecting hate speech, and it utilized a collection of supervised learning algorithms. They combined 5 various classification algorithms, such as K-Nearest Neighbours, Random Forest, Naive Bayes, Support Vector Machine, and Maximum Entropy. They gathered the data set through the Twitter API, and the data set was manually annotated. During the preprocessing, they utilized tokenization, filtering, stemming, and term weighting. They used a bag-of-words feature using TFIDF techniques. The naive Bayes algorithm was the most accurate at 78.3 percent out of the five stand-alone classifiers. P.Sari et al. (2019) suggested a method of hate speech detection with the use of logistic regression on Twitter. They gathered the information on Twitter and used Case Folding, Tokenizing, Filtering, and Stemming techniques in the pre-processing stage. The TF-IDF method is applied to the data after pre-processing to do the vectorization. The Logistic regression algorithm has also been used after Feature engineering, and they have discovered 84% of the accuracy. Oluwafemi et al. (2020) put forward a method of identifying a tweeter's offensive speech. The data set was gathered by the author through the Twitter API and labeled the data set into two categories: free speech (FS) and hate speech (HT). To clean the data, they used special characters, emojis, punctuations, symbols, hashtags, and stopwords in the preprocessing stage. During the feature engineering phase, they used the TF-IDF method to convert the text into feature vectors. They have achieved 89.4 percent of accuracy after using the support optimized vectors machine using n-grams.

Another method to recognize hate speech on Instagram in 2020, Annisa Briliani et al. suggested the application of the k-nearest neighbor classifier. The data set was gathered with the help of the Instagram API of Instagram and annotated those data set manually. They categorized the dataset into 2 labels, i.e., 0 and 1. The data were cleaned during the preprocessing stage, and feature engineering was used with TF-IDF. Subsequently, they used the k- nearest neighbor algorithm and discovered 98.13 percent accuracy. Patti et al. (2019) proposed a hybrid hate speech detection design that combines a Linear Support Vector Classifier (LSVC) with a Long Short-Term Memory (LSTM) neural network using word embeddings. They combined linguistic tools like HurtLex and implemented multilingual word embeddings. Their shared learning model obtained an F1-score of 68.7, and it is evidence that the integration of lexical features with machine learning strategies is effective.

The model of hate speech detection in Arabic social media was created by Alsafari et al. (2020) based on the data gathered on Twitter. The data was grouped under four categories, namely, Religious, Nationality, Gender, and Ethnicity. Once they preprocessed the data, they applied CNN and BERT-based models to the data to obtain an F1-score of 75.51%. Their paper emphasizes that deep learning solutions are effective in detecting hate speech in Arabic. Gamback et al. (2019) used deep learning models that implemented CNN to identify hate speech on Twitter and trained the models to identify data as sexism, racism, combined sexism and racism, and non-hate speech. Their most successful model with feature representations using word2vec and character n-grams got an F1-score of 78.3%. In their study, Sylvia Jaki et al. (2019) adopted an unsupervised method of hate speech detection in Twitter. They gathered over 50,000 tweets and clustered identical biased words with the NLP techniques. Their approach had an F1-score of 84.21, demonstrating that clustering approaches have the capability to find patterns of hate-related language.

Michele Di Capua et al. (2019) have suggested an unmonitored approach to identifying cyberbullying based on the data obtained on YouTube. They collected more than 54,000 comments and hand-annotated them. Their model was accurate at 64 percent using the GHSOM algorithm and 10-fold validation.

Research Methodology

Research Design

In the paper, the research design is an exploratory computational study, which aims to study the automatic hate speech detection in Roman Urdu online posts. Instead of proving a certain hypothesis, the study tries to find linguistic trends that define hate speech in the Pakistani online discourse. Because the Roman Urdu hate speech detection is a fairly understudied field in Computational Linguistics, the exploratory method enables the patterns and structures to be revealed through the data. This is analyzed based on corpus software, the AntConc, which allows systematic study of the lexical items, the frequency, and distribution of such items. This design combines computational text-based processing with discourse-based interpretation in order to gain a better insight into the phenomenon and practical application of hate speech in online communication.

Research Method

The research adopts a mixed-method study involving the use of keyword based automated detection method and the qualitative linguistic analysis. Hate speech is quantitatively detected with a rule-based algorithm with a predefined set of keywords, processed using the corpus tool AntConc to get frequency counts, concordances, and patterns of keywords. Qualitatively, the analysis of the study is based on the interpretive approach, where the Critical Discourse Analysis and Speech Act Theory are used to understand how a choice of comments is constructed to create hostility, discrimination, or antagonism against groups. In this combined method, hate speech is not only examined as a single lexical item but also as a social and discursive activity that has a meaning, which brings together a combination of computational accuracy and theoretical insight.

Theoretical Framework

The paper is based on the Critical Discourse Analysis (CDA) and Speech Act Theory. Critical Discourse Analysis serves as a guideline to the study of the way language builds power relations, ideological positioning, and social dominance in online communication. Using CDA, the study explains the way in which hate speech displays wider socio-political issues, group identities, and discursive practices in Pakistani online environments. The analysis is further facilitated through the Speech Act Theory that looks at the way utterances are actions. Hate speech does not just describe something, but it is a type of speech that does things, like giving insults, threats, degrading or marginalizing some people and groups. Through the use of Speech Act Theory, the paper evaluates the way some of the lexical options in the Roman Urdu remarks serve as performative acts of aggression. Collectively, these structures enable the study to go beyond the superficial level of detecting keywords and understanding the communicative purpose behind the hate utterances.

Corpus Preparation and Data Collection

The dataset in this paper will include 108 Roman Urdu comments that were gathered in the publicly available posts on Twitter (X) in the socio-political environment of Pakistan. The remarks were chosen based on the discussions connected with political discussions, social topics, and controversies in society because these areas have more chances to produce polarized and emotional reactions. The sampling was purposive such that the sample of comments covered a variety of opinions, including possibly hateful and non-hateful remarks.

The data was then collected, cleaned, and normalized systematically. Hyperlinks, hashtags, user mentions, emojis, and duplicate entries were eliminated, and the text was fixed to have consistency. Minor normalization was also done since Roman Urdu does not have standardized spellings, so that the spelling variation of frequent abusive words or appraisive words should be minimized. The filtered-out comments were collected and analyzed as a plain text corpus.

An index of hate speech keywords lexicon was subsequently developed using words with frequently occurring derogatory, abusive, and discriminatory terms that were recognized in the initial reading of the data. In AntConc, the rule-based detection was based on this lexicon. The completed corpus was submitted to AntConc for frequency analysis and identification of keywords as a pattern.

Data Analysis

The corpus that is to be analyzed is a collection of 108 online comments posted on Twitter (now X), written in Roman Urdu, the Romanised form of the Urdu language, Latin scripts, mostly that is used in digital communication in Pakistan. The remarks are mostly aimed at high-profile political leaders, such as Chief Minister Maryam Nawaz, Prime Minister Shehbaz Sharif, former Prime Minister Imran Khan, Bilawal Bhutto Zardari, a few of the prominent journalists of Pakistan, and the institutional machine of the Pakistani military establishment. The data sample is indicative of a highly tense sociolinguistic space, where any kind of public, political speech is intertwined with the issues of gender, religion, ethnicity, nationhood, and authority.

The corpus had to be preprocessed thoroughly before analysis could be done. Roman Urdu has certain methodological issues: no standard orthography, which implies that one and the same lexicon can be represented in a variety of spelling patterns in different comments. An example of this is the slur *kuti,* which is used in different forms in the dataset, namely *kuti, kutti, and kutia*. There are also instances of switches between the Urdu and English languages; English loanwords like *fake* and *puppet* are also inserted in otherwise Urdu sentences. These characteristics demanded systematic normalisation before the building of the lexicon, and the identification of keywords and all decisions of the analysis are recorded in an open manner throughout the sections below.

After the process of labelling using the rules, the 108 comments were coded into three categories. The number of comments that were classified as hate speech due to including explicit slurs, dehumanising animal comparisons, gendered sexual attacks, divine curses against people, or compound identity-based attacks was 21 (19.4%). 45 (41.7%) comments were determined as non-hate speech, which included political criticism, statements of support, prayers, and neutral or observational commentary. The rest of the comments (38.9%), marks that were borderline, characterized by hostile rhetoric, degrading metaphors, conspiracy framing, group-based identity attacks, or moral condemnation, were kept but did not entail the explicitness of an explicit Tier 1 slur. The reason this third category is not collapsed into a binary label is that it is considered to be preserved as an independent unit of analysis because it embodies the real pragmatic ambiguity that cannot be eliminated by the process of keyword detection.

The distribution of comments of the three categories is in itself analytically significant. The largest single category is the borderline category (38.9%, n=42), which is slightly higher than the non-hate speech (41.7%, n=45) and significantly higher than the unambiguous hate speech (19.4%,n=21). This allocation shows that most linguistically violent material in this politically charged corpus falls somewhere within a gray area of hate speech on the one hand and reasonable political critique on the other, which is analytically unseen in purely keyword-driven paradigms and the rationale why the stratified methodology of this study was undertaken.

Lexical Patterns and Keywords in Rule-Based Detection

An inductive inference was made on a hate speech lexicon based on the corpus and used as a detection tool. AntConc's WordList and KeyWord functions were used to determine keywords and rank them. The lexicon that resulted was categorized into three levels based on the level of semantic explicitness and depending on the context-dependence of the hate-marking.

Tier	Term (Roman Urdu)	English Gloss	Freq.	% HS (n=21)	Semantic Domain
Tier 1	*kuta / kuti / kutto*	Dog (m./f./pl.)	10	47.6%	Animal dehumanisation
Tier 1	*harami / harm**	Bastard (m./f.)	10	47.6%	Lineage slurring
Tier 1	*haramkhor*	One who eats ill-gotten gains	1	4.8%	Moral-lineage slur

Tier	Term (Roman Urdu)	English Gloss	Freq.	% HS (n=21)	Semantic Domain
Tier 1	*rakhail*	Kept woman / mistress	1	4.8%	Gendered sexual slur
Tier 1	*kanjar khana*	Brothel / den of vice	1	4.8%	Gendered sexual slur
Tier 1	*jahannami*	Destined for hell	1	4.8%	Religious condemnation
Tier 1	*khusra*	Eunuch (transphobic slur)	1	4.8%	Transphobic identity attack
Tier 1	*nasali / nasal*	Racial / of bad lineage	2	9.5%	Lineage / racial slur
Tier 1	*ghaddar*	Traitor	2	9.5%	Nationalist betrayal
Tier 1	*badmash*	Criminal / thug	1	4.8%	Criminal labelling
Tier 1	*lomri*	Fox (animal metaphor)	1	4.8%	Animal dehumanisation
Tier 1	*pagal kutta*	Mad dog	2	9.5%	Animal dehumanisation
Tier 1	*Chuchra / Billo*	Scoundrel / tom-cat	2	9.5%	Animal / identity slur
Tier 1	*phuddo*	Fool/idiot (vulgar)	1	4.8%	Intellectual dehumanisation
Tier 1	*chomu*	Fool / idiot (derogatory)	1	4.8%	Intellectual dehumanisation
Tier 2	*bayghairat*	Without honour / shameless	8	38.1%	Honour-shame discourse
Tier 2	*besharam*	Shameless	5	23.8%	Honour-shame discourse
Tier 2	*munafiq*	Hypocrite	6	28.6%	Religious-political labelling
Tier 2	*jhootha / jhooti*	Liar (m./f.)	7	33.3%	Credibility attack

Tier	Term (Roman Urdu)	English Gloss	Freq.	% HS (n=21)	Semantic Domain
Tier 2	*ghatiya / ghatia*	Base / low / cheap	3	14.3%	Social degradation
Tier 2	*dogla*	Two-faced / hypocrite	2	9.5%	Character attack
Tier 2	*makar*	Cunning / deceitful	2	9.5%	Character attack
Tier 2	*Patwari*	PMLN supporter (derogatory)	1	—	Group identity slur (BL)
Tier 3	*Allah ka azaab*	God's punishment	3	14.3%	Divine imprecation
Tier 3	*Khuda ki lanat*	God's curse	2	9.5%	Divine imprecation
Tier 3	*yeh yaad rakhna*	Remember this	2	9.5%	Veiled threat/warning
Tier 3	*hisaab zaroori*	Reckoning is necessary	1	4.8%	Implicit threat
Tier 3	*parcha kato*	File a complaint	1	4.8%	Incitement directive

Table:1 Frequency of Hate Speech detected keywords

Tier One: Primary Slurs and Explicit Profanity

The first level is made up of lexical items that serve to instigate an unambiguous slur in practically all settings. They are degrading words, the main semantic operation of which is to humiliate. The most common and statistically salient words of this type are the words that refer to lineage and moral legitimacy. The bare two-word comment *Kuti haramzadi* (Comment 16) piles two Tier One slurs, a gendered animal slur, and a feminine lineage slur, in what is the narrowest occurrence of hate speech in the corpus, and with nothing but abuse in it. Likewise, *Harami naslain* (Comment 67) combines the ancestry reproach *harami* with naslain (generations/bloodlines), and spreads the disapproval over a whole kin group. They have high values in the AntConc analysis, and these are *harami* and its derivatives.

Animal slurs represent the other domineering sub-cluster of Tier One, which is the statistically strongest category of hate indicator in the dataset. The masculine *kutta* (dog) and *kutto* (plural of kutta) are found in Comment 54, *Fauj ky khilaf bakwas krne waly kutto* (dogs who speak nonsense against the army), when used not against an individual, but on a crowd, as a dehumanising group name to describe a political stand. In Comment 99, the animal metaphor is most elaborated, using the compound *pagal kutta* (mad dog) twice in a single utterance and then continuing the metaphor into the predicate, *bonkta he kutta* (the dog barks), to make the speech acts of the target coded as sub-human noise.

Yar ye iqrar he ya Kya he Yar ye to bilkul Pagal kutta jaisa he iska pass 2sra koi kamm nhii he Sara dinn tweets krta he or Imran Khan or uska behno pe bonkta he kutta ye aramii ka baCha k pass 2sra koi kamm nhii.

The third sub-cluster is gendered sexual slurring. Comment 10 has (prostitutes in English) as an object of a compound sentence that also contains the word *chaklay* (brothels) to sexualise the political context of the target, whereas Comment 17 refers to *rakhail* (kept woman/mistress) as an object denouncing the character of a named person. The two terms are only used to address political women in this corpus. At the Tier One level, religious denunciation takes the form of *jahannami aurat* (Comment 33), a dual word phrase that combines spiritual condemnation that lasts forever with gender address, and a curse that is God-based *Khuda ki lanat ho un par* (may God's curse be upon them, Comment 35), which is the culmination of a conspiracy-based assault.

Tier Two: Context-Dependent Derogatory Descriptors

The second level includes lexical elements whose hate-marking possibility is factual but contextual. *Bayghairat* (without honour), *munafiq* (hypocrite), *jhootha / jhooti* (liar), *ghatiya* (base/low), and *besharam* (shameless) are among those words that are very common in the hate speech category as well as in the borderline category of the corpus. The difference between their application in hate speech remarks and their application in borderline remarks lies in the directionality of the attack. When used on conduct or policy, it serves as criticism; when used on the very aspect of the identity of an individual, to what he has done, it serves as degradation. This is illustrated in the comment Besharam bayghairat sahafi (Comment 71): There are two Tier Two terms, which are in direct apposition to the address with no policy referent, and the identity attack is the only communicative action of the utterance.

Moreover, in the borderline category, Tier Two markers are represented in more contextualised types. Politically internalized slur *Patwari* (Comment 18), used in reference to PMLN supporters but meaning commentators, works as a group identity signifier, thus focusing on political affiliation instead of personal character. Unlike a Tier One slur, it is semantically less explicit, but its role is similar: it identifies a group of people as worthy of being removed from the life of the population (*nikaal kar dekho,* attempt to remove them) with the help of an embarrassing term. The case demonstrates the role of the sociolinguistic context in the context of hate speech classification: words that do not reveal themselves in a keyword lexicon can have substantial derogatory social implications in their community of use.

The most common word in Tier Two is the *bayghairat*, which is used in Comments 36, 40, 51, 60, and 71. It is powerful because the centrality of honour (ghairat) is a component of South Asian systems of social values: to describe a person *bayghairat* is not only to denounce their actions but also to announce that they do not belong to the moral community. Likewise, *munafiq*, which appears in Comments 19, 21, 94, and 100, has some special weight in an Islamic cultural setting, where the Quranic character of the hypocrite is deemed to be morally inferior to that of an unbeliever. An example of Tier Two terms used to create ironic delegitimisation by means of official titles is the *jhooton ki wazir-e-ala* (Comment 32, Chief Minister of liars).

Tier Three: Escalatory and Threatening Language

The third level is made up of utterances that indicate rhetorical extremism, but do not necessarily include direct slurs. Their hate-marking capacity consists in their illocutionary force, which consists in what they do and not what they actually say. The comment 34, *Tum logon par Allah Ta azaaba ka azaab nazil ho ga yeh yaad rakhna* (the punishment of God Almighty will fall on you all, remember this) uses an imprecation of God, heinous and a directive imprecation (threat), which has a dual illocutionary effect: commissive (curse) and directive (threat). The most institutionally decisive example in this tier is comment 50: it is used specifically against named persons, and specifically against a military Corps Commander it demands that the targets should face formal punitive action, *seedha parcha kato* (file a complaint directly), i.e., verbal condemnation should give way to one that is demanded by the state. This remark is not a slur, but is analytically at the top of the corpus as such, solely because its perlocutionary ambition has gone beyond language to an institutional action.

Linguistic Patterns, Collocations, and Recurrent Structures

AntConc collocation process was used to extract collocations. The node *besharam* (shameless) has the most significant collocation with the word *aurat* (woman) at position to form the compound *besharam aurat*, a gendered shame attack which is repeated throughout the corpus and is the most salient hate-marked multi-word unit in the dataset. It is also found in collocating with *bayghairat* in the neighbouring terms in terms of constructing a reinforcing honour-shame cluster where the two terms enhance each other. The deictic pronouns *yeh* and *iss* (this / this one) are regularly observed in the corpus, and they do not serve as neutral pointing, but rather mark the target as close and yet ethically unrelated.

The most notable collocate around the node *kuta / kuti* is the neighboring *harami*, which forms the word *harami kutta / kutty*, or the sum of lineage slur and animal slur, which is most frequently used when directed at perceived political traitors. In Comment 84, this tendency is pushed to its utmost extreme, where five abusive words are piled up cumulatively against two named politicians.

Such a statement performs what can be termed as cumulative dehumanisation. The successive term enhances the previous ones, and the targets named are framed with an abuse that leaves no semantic space to dignify counter-representation. The node *jhootha / jhooti* has the most salient collocate *munafiq*, which validates the hypothesis that the credibility attacks and hypocrite-labelling are a consistent compound strategy. An example of this trend can be found in Comment 19, *jhooti munafiq aurat* (lying hypocrite woman), where the two words are used in series on a female politician, and their combination creates an identity attack, both gendered, moral, and religious in register at the same time.

Recurrent Syntactic Patterns

The concordance analysis of hate-labelled and borderline subsets (n=63) shows that there are five common syntactic templates with the help of which abusive meaning is systematically formed in this corpus. The most widespread and the most used is the deictic-plus-slur-plus-target form, where a proximal deictic (yeh / iss) comes before a slur or a derogative marker, which is in turn directed towards the target. The othering and identity attack in this structure is undertaken together in the simplest syntactic frame: *Iss besharam aurat ko nahi sharam hai nahi haya.*

The parallelism nahi ... nahi (neither ... nor) only heightens the reproach, so that the lack of virtue is complete. The second strategy is ironic title inversion, whereby a formal title is grammatically reworded to take away its legitimacy. The *jhooton ki wazir-e-ala* (Chief Minister of liars, comment 32) turns the dignity of the office in on him and replaces it as a collective stigma. The third one is the imperative expulsion order, which is represented by the speech acts like *Chal duffa ho* (Comment 19, get lost) and *Bakwas band kar istifa do* (Comment 63, stop talking nonsense and resign), which require the target to be pushed out of the social space. The fourth is the divine condemnation formula, which is a structurally rigid phrase-type of the sort [divine agent] + [punishment/curse term] + [target or second person], which has been attested in Comments 34 and 35. The fifth, which is already depicted above, is cumulative apposition where the several slurs are piled together without connectives to result in intensified dehumanisation.

Multi-Word Expressions

Many phrases in multi-words serve as fixed formulae in the corpus. The compound *besharam aurat* (shameless woman) is most commonly used and is always directed against female politicians. Another interesting example is the animal-metaphor compound *pagal kutta* (mad dog, comment 99), which is uniquely extended into the predicate by the verb *bhonkna* (to bark), and which therefore represents a dehumanising metaphor of the communicative behaviour of the target as well as his or her identity. *Allah ka azaab*, the religious imprecation used as a formulaic opening of the condemnatory utterances in Comments 34 and 35, and *Khuda ki lanat*, the imprecatory utterance used as a formulaic closer of Comment 35. Targets are delegitimised through the conspiracy-frame compound *America ka puppet* (Comments 35, 36), which places them in the role of foreign agents against the Muslim community. In the comment, *Aur jab jab corruption ka naam koi sunay ga to uske zehan mein yeh......* (joined at the hip Comment 10) carries abusive and offensive innuendo, trying to reclassify a political party as a criminal organization, which is a declarative speech act in the taxonomy of Searle, the categorical label *dehshat gard giroh* (terrorist group, comment 23).

How Online Comments Enact Hate Speech: CDA and Speech Act Theory

Speech Act Classification

Of the 21 hate-labeled comments, there are 5 types of illocutionary acts. The dominant one is the expressive utterances, which are an outlet of the emotional attitude of the speaker towards the target that are not of a chief directive nature. The most frequent types of hate speech acts in the corpus are expressives, which involve most of the Tier One slurring: such as comments *like Kuti haramzadi* (Comment 16), Jahannami aurat (Comment 33), and *Harami naslain* (Comment 67), which do not ask anyone for anything in particular; they are engaging in contempt in its purest form.

The second most important type of hate acts is the directive ones, and they are analytically significant in terms of the fact that they shift the hate speech to the action. The indirect consequence of the hate speech act is that the intended perlocutionary act, a formal penalty against explicitly named individuals under a military institution, has broader implications and scope than the utterance itself, as reflected in comment 50, *seedha parcha kato Cor Commander Peshawar kay naam* (file a complaint directly, addressed to the Corps Commander Peshawar). Comment 19 is a mixture of expressive and directive: *Fake minister tujhe kya lagta hai teray jhooti munafiq aurat ki baat pay koi yaqeen karay ga? Chal duffa ho* starts with a delegitimising expressive frame and finishes with an expulsion command.

In the context of Searle, declarative acts are statements that strive to affect a social reality change by being said. When Comment 23 identifies a political party as *dehshat gard giroh* (terrorist group), or Comment 32 identifies a sitting Chief Minister as j*hooton ki wazir-e-ala* (Chief Minister of liars), the speaker is not saying the target but reclassifying, trying to deprive the target of legitimacy by giving him the name of a terrorist group or liar. The commissive acts, which are symbolized by the Comments 34 and 35, involve the use of divine vengeance, namely *Allah ka* *azaab nazil ho ga, Khuda ki lanat ho un par,* where the speaker acts as a channeled conduit of God's judgment and not as one responsible for the threat.

Speech Act Type	**Definition (Searle, 1969)**	**Corpus Examples**	**n**	**%**
Expressive	Expresses the speaker's contempt or hatred toward the target	*Kuti haram*; Jahannami aurat; Harami naslain*	10	47.6%
Directive	Commands target to act or mobilise the audience against the target	*Chal duffa ho; Seedha parcha kato; Bakwas band kar istifa do*	5	23.8%
Declarative	Reclassifies the target's identity or institutional status	*Dehshat gard giroh hai; Jhooton ki wazir-e-ala*	3	14.3%
Commissive	Commits divine agent to punishment; imprecation or veiled threat	*Allah ka azaab nazil ho ga; Khuda ki lanat ho un par*	2	9.5%
Assertive	Factual accusation providing ideological framing for hate speech	*America say paisay liay hain; Corruption kar gayi*	1	4.8%

Table no 2: Distribution of Illocutionary Speech Act Types Across Hate Speech Comments

Gendered Honour-Shame Discourse

The most intensive discursive circulation within the corpus is that of gendered honour-shame defining female politicians, most consistently Maryam Nawaz. This trend builds upon the strongly embedded South Asian patriarchal beliefs that the legitimate social status of a woman is that of the domestic, the personal; her passage into the realm of political action is encoded as transgression, her power is undermined not by policy criticism but by assaults upon sexual character and feminine decency. All the comments that are involved in this pattern include comments 10, 12, 16, 17, 19, and 33, and they all serve to prove that female politicians in this corpus receive a qualitatively different kind of abuse, which strikes at identity as a woman, rather than as a political opponent.

From the perspective of CDA, comment 10 does not criticize policy but realizes what van Dijk (1993) calls negative other-presentation: the role of the female politician is absorbed into the category of sexual transgression, and the province where she governs is presented as ethically polluted by her existence. The compound *besharam aurat* (Comment 12) is a closer-to-the-point declarative shame-assignment: the parallelism *nahi sharam hai nahi haya* conducts an overall moral audit of the target and declares her as she is lacking all the feminine virtues.

Religious Delegitimization and Divine Imprecation

The second dominant trend is the use of Islamic religious power to denounce political enemies. This approach brings the level of personal hatred to a higher plane, and the targets are not only politically bad but also spiritually sinful and punishable by God. In Comment 35, the researcher observes this tendency at its full architectural will: *America say paisay liay hain ab yeh* *apnay Musalmanon ka khoon kar kay wahan say dollar kamaein gay yeh mulk mein bhi dehshath gardi yahi fauji kara rahay hain, Khuda ki lanat ho un par.*

This comment creates a conspiracy frame in the first clause, and it plays an assertive role, thus offering apparent rational support to the curse in the latter. The utterance is followed by the imprecation *Khuda ki lanat ho un par*, a commissive utterance, which appeals to God as a way of making the condemnation unconditional. As the reading of the CDA shows, religious terms (Musalmanon, Khuda, lanat) can be defined as ideological legitimisation: the appeal to God changes personal hatred into righteous indignation. The phrase munafiq, used throughout the other comments, works in the same fashion, citing the authority of the Quranic signifier to make the issue of political dissent the issue of religious betrayal.

Dehumanization Through Animal Metaphor

One of the most widely studied phenomena in hate speech research is dehumanisation with the use of animal metaphor that is statistically predominant in this corpus (Musolff, 2016). The slurs kuta / kuti / kutto and lomri (fox) are the primary nodes. In comment 31*, Makar khandan ki makar lomri* (a fox who is a fox, of a foxy family), the metaphor of animals is extended to a whole group of kin, naturalising the corrupt and making the political dynasty irredeemable. The word makar in the noun and its genitive modifier is repeated to create a rhetorical echo, which performs what van Dijk (1993) describes as collective negative other-presentation. The most continued subject matter is comment 99, which uses the animal slur twice and then turns it into a verbal metaphor (*bonkna*, to bark) to code the communicative activities of the target as being sub-human.

Deictic Othering and Categorical Labelling

As mentioned in the collocation analysis, the proximal deictic *yeh / iss* is a steady characteristic of hate speech and borderline words in this corpus. As an analysis using CDA shows, it can be seen as a discursive device of othering: by isolating the target with a demonstrative, the speaker places them socially close but morally distant. It is most directly manifested in Comment 12 *(Iss besharam aurat,* this shameless woman) and Comment 21 (*Yeh munafiq chup hai,* this hypocrite is silent) in a process that Wodak (2001) refers to as discursive othering. In Comment 21, the structure *yeh* + categorical label + predication performs a declarative identity assignment in nine words: the target is identified, categorized as hypocrite, and charged with moral dereliction within a single minimal clause.

Incitement Directives and the Perlocutionary Dimension

The last pattern is that of comment, the major illocutionary activity of which is that of directive, namely, calling the audience to action, but not the target. Such remarks shift hate speech out of the expressive register, into the action-oriented register, and such an utterance has the action-oriented element of a perlocutionary act: what an utterance is meant to accomplish to the hearer, in keeping with the concept of the perlocutionary act presented by Austin (1962). The most obvious example in the corpus is Comment 50: *Yahan mazammat say kaam nahi chalay ga balkay seedha parcha kato Cor Commander Peshawar kay naam*

There is no explicit slur in this remark. It is hate speech in all of its directive power and institutionalized target: the commentator does not merely reject verbal protest as insufficient; he insists on official punitive action, which is addressed to a particular military leader and applied to individual persons. The logic of comment 37 is like that of a similar comment: *yeh ghaddar generals judges hain inka hisaab har haal mein zaroori hai* uses the same word ghaddar (traitor), and uses the same word hisaab (reckoning), but includes a veiled threat, functioning as an implicit commissive act.

Discussion

The three-level analysis of the lexical keyword detection, collocation and patterning analysis, and discourse-pragmatic interpretation presented in this study shows that the hate speech in this Roman Urdu Twitter corpus is a patterned, ideologically coherent, and well-structured discursive practice. Various cross-cutting observations are worth highlighting. The most significant observation is the overlapping of gender and political opposition. The type of abuse that female politicians get is qualitatively different and quantitatively more severe than the one received by male politicians. Male politicians are slandered on inheritance, devotion, and position in religion; female politicians are slandered on all these, as well as sexual integrity and female decency. This twofold pressure has been shown on the ideological cross-lacing of the social system of patriarchy and political power reproduction by means of language, and it can be seen in six of the 21 hate speech remarks in the corpus, an overproportionate concentration, seeing as there were more than one or two targets in the corpus.

The allocation of comments by three categories: hate speech (19.4%), borderline (38.9%), and non-hate speech (41.7%) is also analytically significant. The borderline is the most extensive of the three, and almost twice as much as the unambiguous hate speech category. This proves empirically that most linguistically aggressive material in this corpus lies in a grey area between outright hate speech and legitimate political criticism. Other words, like *munafiq, bayghairat, and Patwari*, are also present in the hate speech and borderline subsets, which proves that the frequency of keywords in them cannot be a reliable criterion between hate speech and intense political criticism in Roman Urdu. The focus of the attack, be it on identity or conduct, is the key variable of analysis, and it is one that no keyword list can capture. This conclusion places rule-based detection as not an ultimate classification tool but a primary filtering tool that needs to be followed up by pragmatic and contextual analysis to accomplish its task.

The role of religious language in this corpus is always as a legitimisation strategy of hatred and not as a spiritual expression in itself. Speakers make personal hostility a religious duty by justifying their attacks as punishments of God or traitors to the Muslim world. Here, this specifically Pakistani approach is the process of incorporating the Urdu Islamic vocabulary into otherwise colloquial Roman Urdu registers, a code-mixing practice by which the religious denunciation of the language simultaneously becomes particularized and authoritative.

Conclusion

This data analysis has revealed that hate speech on the Twitter comments of Roman Urdu is an organized, patterned, and ideologically-driven discursive practice. The sample of 108 comments produced 21 examples of hate speech (19.4%), 42 examples of borderline (38.9%), and 45 examples of non-hate speech (41.7%). A three-level lexicon of hate speech was created based on the analysis of AntConc keywords and collocating words, including slurs that refer to five semantic domains: animal dehumanisation, lineage and legitimacy slurring, honour-shame discourse, religious condemnation, and gendered sexual abuse. In Critical Discourse Analysis and Speech Act Theory, it was found that there are five significant discursive patterns: gendered shame discourse, religious imprecation, animal dehumanisation, deictic othering and incitement directives, all of which have similar though not identical social functions of degradation, exclusion and threat.

The most important theoretical contribution of this analysis is that the hate speech in this corpus works both on the surface and the ideological level, where, on the surface, it targets named individuals, whereas, at the ideological level, it reproduces and strengthens the already existing hierarchies of gender, religion, ethnicity, and political affiliation. Moreover, the methodological discovery that Comment 18 is classified hate speech, as a politically entrenched group identity slur, when examined, demonstrates a methodological result with general implications: automatic detection of hate speech in roman Urdu will only be made available by not just a keyword lexicon, but also more profound sociolinguistic understanding of the derogatory vocabulary of community-specific group identities which lies beneath the usual slur lists. Language is never described; always already doing social work as Critical Discourse Analysis demands. Social work in this corpus consists of the ordered production and reproduction of hatred using the materials of the Roman Urdu language.

Future Recommendations

Future studies on this topic should focus on some of the directions that arise due to the shortcomings and the findings of the current research. First, to have the ability to calculate the keyness and collocation scores in AntConc with proper statistics, the creation of a larger and multi-platform Roman Urdu corpus with a formal reference corpus would facilitate the quantitative aspect of rule-based detection. Second, inclusion of machine learning solutions, specifically the fine-tuning of multilingual language models based on labelled Roman Urdu data, would significantly enhance the classification accuracy, especially on the borderline category that took 38.9% of the current corpus and was not easily detected using only keywords. Third, native speaker consultation should be involved in future annotation in order to document politically embedded derogatory words like Patwari, which have a lot of social relevance in the community of use, but cannot be identified by the normal lexicons. Fourth, the overproportionality of female politicians as the victims of gendered hate speech in this corpus should be researched as a specific research line, since this issue has a certain implication on the role of female politicians in the Pakistani state affairs. Lastly, a longitudinal research design of hate speech language over time and political actions would help define whether the lexical patterns and discursive strategies found in this research would be long-term stable aspects of Pakistani digital political culture or transient phenomena that respond to the political situation.

References

Allison Lahnala, Charles Welch, Béla Neuendorf, and Lucie Flek. 2022. Mitigating toxic degeneration with empathetic data: Exploring the relationship between toxicity and empathy. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4926–4938, Seattle, United States. Association for Computational Linguistics.

Alsafari, S., Sadaoui, S., & Mouhoub, M. (2020). Hate and offensive speech detection on Arabic social media. *Online Social Networks and Media*, *19*, 100096.

Anthony, L. (2022). AntConc (Version 4.2.0) [Computer Software]. Waseda University. https://www.laurenceanthony.net/software

Austin, J. L. (1962). How to do things with words. Oxford University Press.

Belavadi, V., Zhou, Y., Kantarcioglu, M., & Thuraisingham, B. M. (2021). Multi-concept adversarial attacks. *arXiv preprint arXiv:2110.10287*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Biagioni R: Sentiment Analysis. *The SenticNet Sentiment Lexicon. Exploring Semantic Richness in Multi-Word Concepts. SpringerBriefs in Cognitive Computation.* Springer; 2016; **4**. : pp. 7–16.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17):14867–14875.

Briliani, A. (2020). Deteksi Ujaran Kebencian Dalam Bahasa Indonesia Pada Kolom Komentar Instagram Dengan Metode Klasifikasi K-Nearest Neighbor.

Chiril, P., Pamungkas, E. W., Benamara, F., Moriceau, V., & Patti, V. (2022). Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, *14*(1), 322-352.

Di Capua, M., Di Nardo, E., & Petrosino, A. (2016, December). Unsupervised cyber bullying detection in social networks. In *2016 23rd International conference on pattern recognition (ICPR)* (pp. 432-437). IEEE.

Fairclough, N. (1992). Discourse and social change. Polity Press.

Gambäck, B., & Sikdar, U. K. (2017, August). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online* (pp. 85-90).

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215-230.

Harris, C., Halevy, M., Howard, A., Bruckman, A., & Yang, D. (2022, June). Exploring the role of grammar and word choice in bias toward African American English (AAE) in hate speech classification. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 789-798).

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*, 126232.

Jaki, S., & De Smedt, T. (2019). Right-wing German hate speech on Twitter: Analysis and automatic detection. *arXiv preprint arXiv:1910.07518*.

Kennedy, B., Jin, X., Davani, A. M., Dehghani, M., & Ren, X. (2020, July). Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5435-5442).

Kotarcic, A., Hangartner, D., Gilardi, F., Kurer, S., & Donnay, K. (2022, December). Human-in-the-loop hate speech classification in a multilingual context. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 7414-7442).

Kunal Relia, Zhengyi Li, Stephanie H Cook, and Rumi Chunara. 2019. Race, ethnicity and national originbased discrimination in social media and hate crimes across 100 us cities. In Proceedings of the International AAAI Conference on Web and Social Media, volume 13, pages 417–427.

Kwok, I., & Wang, Y. (2013, June). Locate the hate: Detecting tweets against blacks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 27, No. 1, pp. 1621-1622).

Lahnala, A., Varadarajan, V., Flek, L., Schwartz, H. A., & Boyd, R. L. (2025, June). Unifying the extremes: Developing a unified model for detecting and predicting extremist traits and radicalization. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 19, pp. 1051-1067).

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021, May). Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 17, pp. 14867-14875).

Musolff, A. (2016). Political metaphor analysis: Discourse and scenarios. Bloomsbury Academic.

Ramponi, A., Testa, B., Tonelli, S., & Jezek, E. (2022). Addressing religious hate online: from taxonomy creation to automated detection. *PeerJ Computer Science*, *8*, e1128.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020, July). Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5477-5490).

Schmidt A, Wiegand M: A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain, Association for Computational Linguistics.* 2019; pp. 1–10.

Searle, J. R. (1969). Speech acts: An essay in the philosophy of language. Cambridge University Press.

van Dijk, T. A. (1993). Elite discourse and racism. Sage Publications.

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S. A., & Margetts, H. (2019, August). Challenges and frontiers in abusive content detection. In *Proceedings of the third workshop on abusive language online* (pp. 80-93).

Wodak, R. (2001). The discourse-historical approach. In R. Wodak & M. Meyer (Eds.), Methods of critical discourse analysis (pp. 63–94). Sage Publications.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3143–3155, Online. Association for Computational Linguistics.

Fatahillah, N. R., Suryati, P., & Haryawan, C. (2017, November). Implementation of Naive Bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech. In *2017 International Conference on Sustainable Information Engineering and Technology (SIET)* (pp. 128-131). IEEE.

Fauzi, M. A., & Yuniarti, A. (2018). Ensemble method for indonesian twitter hate speech detection. *Indonesian Journal of Electrical Engineering and Computer Science*, *11*(1), 294-299.

Ginting, P. S. B., Irawan, B., & Setianingsih, C. (2019, November). Hate speech detection on twitter using multinomial logistic regression classification method. In *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)* (pp. 105-111). IEEE.

Oriola, O., & Kotzé, E. (2020). Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. *IEEE Access*, *8*, 21496-21509.